University of Alberta

# Corpus-Based Learning for Pronominal Anaphora Resolution

by

Shane Anthony Bergsma

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Master of Science**.

Department of Computing Science

Edmonton, Alberta

Fall 2005

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

0-494-09124-X

*Your file Votre référence*
*ISBN:*
*Our file Notre référence*
*ISBN:*

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

---

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

# Abstract

Anaphora resolution is a challenging and important problem in Natural Language Processing. We use machine learning from both labelled and unlabelled corpora to gather probabilistic information for reference resolution. Our unsupervised, unlabelled textual extraction approaches are a form of "bootstrapping" for information extraction. By assuming coreference links in unlabelled text, we can infer statistically meaningful information to assist coreference determination. This includes information on a noun's gender and number, its frequency as an antecedent, and the likelihood of coreference occurring between entities along a given syntactic relationship. These new sources of information are combined with well known constraints and preferences by inducing classifiers using supervised learning.

# Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada, iCORE in Alberta, and a Walter Johns Graduate Fellowship from the University of Alberta.

I would like to thank Dekang Lin for his continued assistance in my work and confidence in my abilities. Dekang is an excellent supervisor; he has laid a strong practical and theoretical foundation for our group's research activities, and he has provided me with opportunities and advantages that I am grateful for every day. I look forward to continued work with him in the future. Also, thank you to Greg Kondrak and the Natural Language Processing Group for providing helpful feedback and criticisms, especially to Chris Pinchak for his daily tips and debugging help. Also special thanks to Colin Cherry for taking a real interest in my research area, guiding me through our project together, and raising the quality of my research practices in general. Thanks to Chris Westbury and my committee members for their time and attention.

Thank you to my parents, Bill and Helen Bergsma, for providing complete and unrelenting encouragement and support throughout my education. Dad has always been a role model for hard work and ingenuity, and helped me find my first technical position at MDS Nordion. Mom fostered a love of reading and of discovering other cultures and languages, both of which continue to motivate me every day. Thanks to Paul and Leanne Bergsma, Susan, Lorne and Kevin Musselman, Cory Dean, Joshua Black, Alex Ferguson, Matt Bulter and Scott Vrooman for their continued support and guidance.

Thank you to Kristin Musselman – more than just my collaborator, mentor, best friend, and lover, but a real part of me and all my work. She deserves at the very least co-author status on this thesis, but will probably settle for my promise of lifelong devotion and attention.

# Table of Contents


**1 Introduction** 1
- 1.1 Coreference and Anaphora Resolution . . . 1
- 1.2 The Anaphora Resolution Process . . . 2
- 1.3 Motivation . . . 2
- 1.4 Our Approach . . . 3
  - 1.4.1 Extracting Information from Unlabelled Text . . . 4
  - 1.4.2 Assuming Coreference Relations . . . 6
  - 1.4.3 Unsupervised Learning of Anaphora Resolution . . . 6
  - 1.4.4 Supervised Learning of Anaphora Resolution . . . 6
  - 1.4.5 Text Analysis . . . 7
- 1.5 Outline of Thesis . . . 7

**2 Related Work** 9
- 2.1 Government and Binding Theory . . . 9
- 2.2 Early Computational Work . . . 12
  - 2.2.1 Evaluation Issues in Anaphora Resolution . . . 15
- 2.3 Machine Learning Approaches . . . 16
  - 2.3.1 Machine Learning for Anaphora Resolution . . . 16
  - 2.3.2 Previous Approaches . . . 17
- 2.4 Bootstrapping . . . 18

**3 Data Sets** 20
- 3.1 Annotated Data . . . 20
  - 3.1.1 Annotation Notes . . . 20
  - 3.1.2 ANC Data . . . 21
  - 3.1.3 AQUAINT Data . . . 22
- 3.2 Unlabelled Data . . . 23

**4 Extracting Gender and Number Information** 24
- 4.1 Introduction . . . 24
- 4.2 Gender in Anaphora Resolution . . . 25
- 4.3 Gender Information Sources . . . 27
  - 4.3.1 Pattern Matching for Noun-Pronoun Gender Pairs . . . 27
  - 4.3.2 Combining Gender Information . . . 28
- 4.4 Experimental Details . . . 29
- 4.5 Noun Gender Classification . . . 30
- 4.6 Anaphora Resolution with Learned Gender Information . . . 32
- 4.7 Conclusion . . . 34

**5 Learning Antireflexive Coreference Relations** 36
- 5.1 Introduction . . . 36
- 5.2 Antireflexive Relations . . . 37
- 5.3 Antireflexive Relation Extraction Algorithm . . . 38
- 5.4 Antireflexive Possessive Pronoun Relations . . . 39
  - 5.4.1 Preposition-Path Filters . . . 39
  - 5.4.2 Verb-Path Filters . . . 41
- 5.5 Experimental Results . . . 42
- 5.6 Conclusion . . . 44

**6 Non-Anaphoric Pronouns** 45
6.1 Pleonastic Pronouns . . . 46
6.1.1 Previous Approaches . . . 46
6.1.2 Our Approach . . . 49
6.2 Cataphora . . . 50
6.2.1 Introduction . . . 50
6.2.2 Cataphora Resolution . . . 51
6.2.3 Scope and Related Work . . . 52
6.2.4 A Cataphora Resolution Strategy . . . 54
6.2.5 Pronoun Compatibility Filters . . . 56
6.2.6 Experimental Results . . . 57
6.2.7 Conclusion . . . 58

**7 Unsupervised Anaphora Resolution** 59
7.1 Introduction . . . 59
7.2 Resolution Framework . . . 61
7.3 Expectation Maximization Anaphora Resolution Learning . . . 62
7.4 Gender Model Initializer . . . 64
7.5 Maximum Entropy Formulation . . . 65
7.6 Experimental Set-up . . . 65
7.7 Results . . . 66
7.7.1 Comparison to Chapter 4 System . . . 69
7.7.2 Top-$n$ Answers . . . 69
7.8 Conclusion . . . 70

**8 Conclusion** 71
8.1 Contributions . . . 71
8.2 Future Work . . . 72

**Bibliography** 74

# List of Tables


4.1 Noun Gender Classification Performance (%) . . . . . 30
4.2 Miscellaneous Micro-averaged Noun Gender Classification Performance (%) . . . 31
4.3 Simple Anaphora Resolution System Accuracy . . . . . 32
4.4 Machine Learned Anaphora Resolution Accuracy . . . . . 33

5.1 Most frequent antireflexive preposition-noun relations (greater than 75% proportion pronouns non-coreferent). Example non-coreferent entities italicized. . . . . 40
5.2 Most frequent antireflexive verb-noun relations (greater than 75% proportion pronouns non-coreferent). Example non-coreferent entities italicized. . . . . 41
5.3 Results for Antireflexive Paths. . . . . 43
5.4 Results for Similar Antireflexive Paths. . . . . 43

6.1 Taxonomy of definite noun phrases . . . . . 48
6.2 Pleonastic detection confusion matrix . . . . . 50
6.3 Results of Cataphora Resolution Test. . . . . 57
6.4 Results of Extended Possessive Filter Test. . . . . 58

7.1 Maximum Entropy Feature Function Weighting. . . . . 67
7.2 Resolution Performance on No-Quotation Test Set (%). . . . . 67
7.3 Resolution Performance on All Pronous, Pronouns from Sentences with Quotations 68
7.4 Comparison to SVM. . . . . 69
7.5 Top-$n$ accuracy on "All" for $n = 1 \ldots 3$. . . . . 70

# List of Figures


1.1 The coreference information extraction process. . . . . 4
1.2 Incorporating coreference information into anaphora classification. . . . . 5
1.3 Example Dependency Tree. . . . . 7

2.1 Example feature vector creation. . . . . 16

4.1 Gender information extraction process. . . . . 25
4.2 SVM anaphora resolution learning. . . . . 33
4.3 SVM anaphora resolution classification. . . . . 34

5.1 Example Dependency Tree. . . . . 38

6.1 Example Cataphora Parse. . . . . 54

7.1 Pronoun resolution instances from Sentence 1. . . . . 61

# Chapter 1

# Introduction

## 1.1 Coreference and Anaphora Resolution

Coreference resolution is the process of determining which expressions in text refer to the same real-world entity. In each of the following examples, we would like a coreference resolution system to determine that the italicized expressions are coreferent:

(1) *John Smith*, current *CEO* of SmithCo, announced *his* retirement yesterday.

(2) *International Business Machines* reported fourth quarter earnings yesterday. *IBM* noted revenue rose by 8%.

(3) Mary has a red *Oldsmobile*. Linda has a green *one*.

(4) *George Bush* was briefly in attendance. The *president* stopped by while on a campaign tour of Ohio.

(5) When the *president* entered the arena with *his* family, *he* was serenaded by a mariachi band.

Anaphora resolution is the important yet challenging subset of coreference resolution where a system attempts to determine which previous entity (the antecedent) a given noun phrase (the anaphor) refers to. While coreference resolution partitions all the nouns in the text into coreferent groups, anaphora resolution seeks preceding antecedents for particular instances of anaphora. Like most researchers in anaphora resolution, we focus on resolving third-person pronominal anaphora (e.g. *he, she, it,* or *they*). Our approaches also handle reflexives (*himself, herself, itself,* and *themselves*).

In the above examples, our anaphora resolution systems will determine the antecedents for the pronoun *his* in Sentence 1 and the pronouns *his* and *he* in Sentence 5.

Although humans can make these resolutions easily, automatic anaphora resolution, and general coreference resolution, remain difficult computational tasks. At the same time, there are many promising research directions within this field. Our work, a synthesis of linguistic insight and machine learning algorithms, proposes several promising new directions as well.

## 1.2 The Anaphora Resolution Process

Computational approaches to anaphora resolution require various natural language processing modules. First, the text must be tokenized, then tagged or parsed. This is needed primarily to identify the noun phrases in the document. Whether the resolution algorithm ultimately works with the flat tagged sentence or leverages the syntax in the parse tree, noun phrase identification is needed to provide the candidates for resolution.

Usually, antecedent candidates are sought in preceding segments of the text. Thus, in Sentence 5 above, our preprocessing steps should provide "president" and "arena" as antecedent candidates to the first pronoun, "his."

When resolving a given pronoun, the system then uses some combination of constraints and preferences (sometimes called factors or features [52]) to identify the correct antecedent noun phrase from the list of potential candidates. At times, more than one noun phrase in the list is an antecedent. We say a system has successfully resolved the pronoun if any one of the correct antecedents has been selected.

In general, the term "constraint" refers to those decision parameters which eliminate possible candidates by virtue of gender and number disagreement, grammar violations, etc. "Preferences" refer to parameters which encourage selection of antecedents that are more recent, more frequent, etc. Implementation of constraints and preferences can be based on empirical insight [38, 35], or machine learning from a reference-annotated corpus [21]. We follow machine learning approaches throughout our research.

In the remainder of this chapter, we motivate the problem, summarize our approach, and outline the breakdown of the paper.

## 1.3 Motivation

Performing anaphora resolution has long been considered a challenging yet vital task for a number of Natural Language Processing applications. To most systems, pronouns and other anaphora are quite uninformative on their own. Anaphora resolution is essentially a way to insert the information that these pronominal symbols represent, and to provide the links to other related phrases and expressions in text. In fact, one can argue that any text-driven system, whether one that automatically clusters similar words [41], discovers inference rules for question answering [44], or determines word translations from bi-texts [10], would have access to more data, determine more relationships, and improve their models by preprocessing their texts with coreference annotation. Anaphora resolution makes texts richer in information.

Enriching texts with more information has a particular benefit in the area of information extraction. Resolving a pronoun to a noun phrase provides a new interpretation for a given sentence, perhaps now revealing something of interest to a Question Answering (QA) system that would have

previously been overlooked. Of course, given that current approaches to anaphora resolution are far from perfect, questions remain about what kinds of resolutions and what level of precision to employ when performing reference resolution for QA [70].

Text Summarization could also benefit from resolution of anaphora. One simple but potentially effective method of Text Summarization is to summarize a document by extracting the introduction, conclusion, and first or last sentences of each intervening paragraph (as done in [68]). Unfortunately, the summary produced by such a method is often incoherent, primarily because pronouns and anaphora are used liberally in these sentences, and the reader of the summary lacks the contextual sentences needed to perform the coreference resolutions [72]. Resolving the pronouns in these sentences before extracting them could provide the information needed for a coherent summary.

We mentioned how anaphora resolution could provide more data to translation algorithms; there are other benefits to anaphora resolution in the area of machine translation. For example, consider the issue of gendered nouns in various languages. In English, we might say "the dog likes its food" and "the cat likes its food," but when we translate the pronoun "its" to German, we need to know whether "its" refers to either a dog or to a cat in order to select the correct gender of pronoun in the translation. We need the masculine possessive pronoun for "der Hund" ("dog" is masculine) and the feminine possessive pronoun for "die Kätze" ("cat" is feminine). Pronoun resolution provides this information.

It should be noted that there are many other issues involved in the translation of pronouns between languages, aside from difference in genders. Sometimes the translation of a pronoun is not another pronoun in the target language at all, but the antecedent itself [54]. In some languages, personal pronouns are left out of dialogue altogether, implied by the conjugation of the finite verb in so-called pro-drop languages [23]. These implicit pronouns are also known as zero-pronouns, and methods have been devised to resolve and translate them [64, 29]. Regardless of the unique challenges of translating between languages, pronoun resolution remains an important component.

Finally, because humans perform anaphora resolution with ease, devising effective methods for anaphora resolution provides a window into human understanding. As we develop better models of text, more intelligent algorithms, and more complete repositories of so-called "world knowledge" to apply to the anaphora resolution task, we come closer to understanding the essence of human dialogue, language and learning.

## 1.4 Our Approach

This thesis details several novel contributions to the field of anaphora resolution, which we outline in this section. First of all, we discuss our methods for gathering information for anaphora resolution from unlabelled text. There are two components to this process: assuming coreference links in text, and then gathering information from these links. How one formulates the assumptions depends on the information to be extracted. We give a flavour of this issue in the following subsections. We then

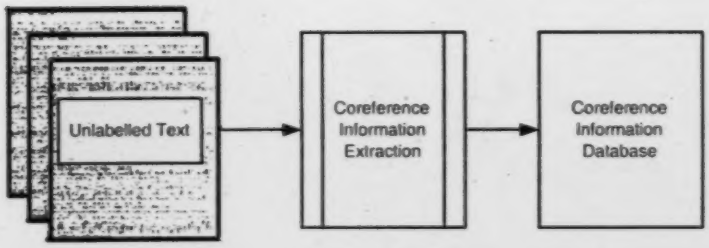



Figure 1.1: The coreference information extraction process.

outline our unsupervised approach to anaphora resolution, where the assumptions and information extraction are done together iteratively. Finally, we highlight the supervised machine learning and text analysis framework used throughout this thesis.

### 1.4.1 Extracting Information from Unlabelled Text

The key contribution of this work is our method for mining useful information from unlabelled corpora to assist anaphora resolution. We use simple heuristics to assume coreference links in unlabelled text, enabling bootstrapping of general coreference resolution information. Each time a link is assumed, some information is collected from the individual instances. In the end, we look at the aggregate data to infer statistically meaningful knowledge.

Figures 1.1 and 1.2 illustrate the general approach. We first apply our extraction algorithms to unlabelled text (Figure 1.1). Then this information, stored in a database, is incorporated into the classification of anaphora, enabling labelling of coreference links in previously unlabelled text (Figure 1.2). These labels could then be exploited by systems doing machine translation, text summarization, or information retrieval, as outlined above.

For example, we learn lexical gender/number information in this manner. An important question when resolving a pronoun to a given noun is: is this noun likely to have the same gender or number as the pronoun? When we see, "Mary thought she looked nice," and we look for the antecedent of "she," we would like to know the probability "Mary" occurs as a feminine singular pronoun. Would we interpret this sentence differently than "Alex thought she looked nice"? What about "The team thought she looked nice"?

One way to determine the gender/number of noun phrases is to look at the aggregate data we collected by assuming coreference links in unlabelled text. In the millions of coreference links that we assumed, what was the probability this noun was linked to a pronoun of this gender/number? Although there might be a lot of noise polluting the individual instances, the question is not so much whether this empirical distribution approximates the true gender/number probability distribution, but whether we can make useful decisions when resolving anaphora using data from these distributions. We show this to be possible. We demonstrate how probabilistic gender information can

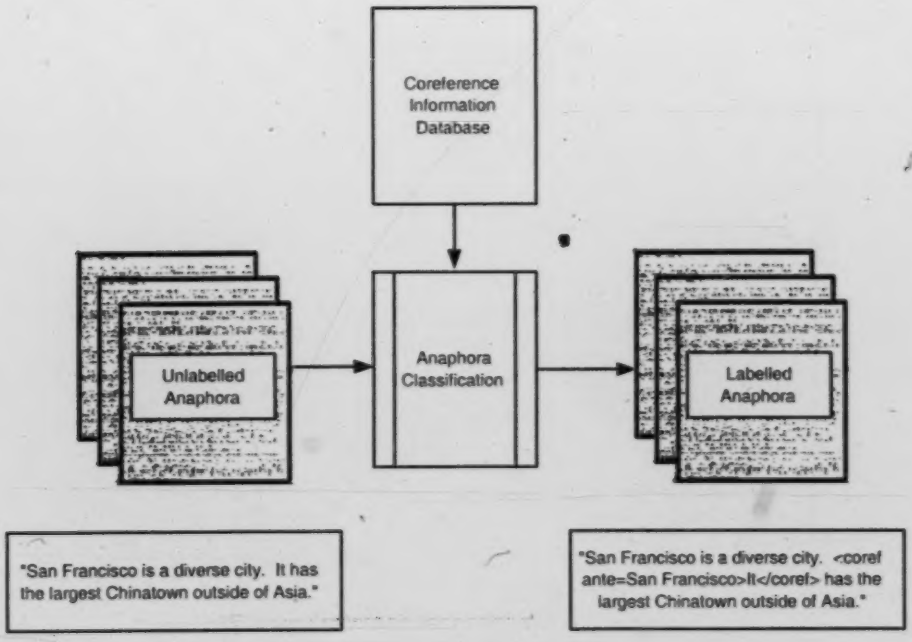



Figure 1.2: Incorporating coreference information into anaphora classification.

boost anaphora resolution performance by over 10% (from 63.2% to 73.3%) over traditional gender constraints.

Another feature learned by bootstrapping simple resolution heuristics is coreference path information. We are able to determine the likelihood of coreference occurring between entities along a given syntactic relationship. The vital question at resolution time is: is this noun likely to refer to the given pronoun along the syntactic relation between the two? How likely are "John" and "his" to refer to each other along the path, "John needs his skis"? Is coreference less likely along the path, "John needs his support"?

We answer this by collecting the paths between millions of same-sentence entities in text, and computing the frequency these entities were either likely coreferent or not coreferent. The given path at resolution time can be looked up to determine likelihood of coreference along the path, and this likelihood can be incorporated into the decision procedure. We show some paths are largely "antireflexive" – meaning coreference is unlikely to occur across them in text. We show this information to be especially useful in resolving possessive pronouns. Resolution of these pronouns is difficult because they do not enable elimination of antecedents using other within-sentence constraints (like the principles of Government & Binding Theory [23], also see Section 2.1).

To demonstrate the utility of this antireflexive path information, we extract phrases in text containing a path between a subject noun and a possessive pronoun. Although resolving these pronouns

to the subject noun is correct in 80% of the cases, our antireflexive path information identifies a significant subset where only 16% of pronouns and nouns corefer.

### 1.4.2 Assuming Coreference Relations

Throughout this paper, we explore different ways to assume the coreference relations in the unlabelled coreference information gathering. For extracting the gender, we use manually-defined lexico-syntactic coreferent paths. We also adopt the novel method of using the unparsed world wide web as a gigantic source of unlabelled data for the extraction of these gender information patterns. For gathering the likelihood of coreference along syntactic paths, we use pronoun-pronoun matches and mismatches along paths in unlabelled text.

### 1.4.3 Unsupervised Learning of Anaphora Resolution

In the final section of this report, we devise a method to gather gender, number, and contextual likelihood information while simultaneously refining the coreference assumption procedure. This is done using Expectation Maximization (EM). By contextual likelihood, we mean: is this noun likely to play the same syntactic role as the pronoun, given the pronoun's parent and its relationship with its parent? Furthermore, we learn: how likely is this noun to even *be* an antecedent?

The process works by counting various occurrences in a sequence of pronoun-candidate list instances, in order to form probability distributions. These distributions are then used to refine the weights in the counting procedure, and the process is repeated iteratively. This Expectation Maximization approach provides a novel unsupervised way of performing anaphora resolution. We not only show that unsupervised learning of anaphora resolution is possible, but we demonstrate levels of performance comparable to supervised approaches.

### 1.4.4 Supervised Learning of Anaphora Resolution

Supervised machine learning is used to combine our resolution constraints and preferences in the two full anaphora resolution systems outlined in this thesis. For our gender information testing, Support Vector Machines are used to select and weight a large set of linguistically-motivated features, including our gender/number probabilities. For testing of our unsupervised approach to anaphora resolution, Maximum Entropy is used to refine the weights on the probability models learned through Expectation Maximization.

Having all layers of our system – the textual extraction of information, the combination of features for classification, etc. – programmed with either unsupervised or supervised machine learning provides us with a system automatically re-configurable for different domains, languages, and dependent technologies.

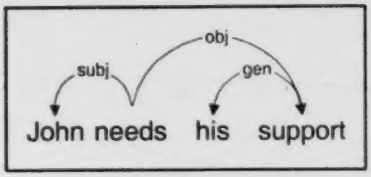



Figure 1.3: Example Dependency Tree.

### 1.4.5 Text Analysis

Throughout this thesis, all text analysis work was done within the framework of Dekang Lin's LaTaT: Language and Text Analysis Tools [43]. The key component of this system is Dekang Lin's minimalist parser Minipar [42]. Minipar is a broad-coverage, dependency-relationship parser, and a descendant of Principar [39]. In a dependency-based parser, each word, called a modifier, modifies at most one parent, its head, although a head can have more than one modifier. Normally a tree is induced from the dependency relationships, with the root of the tree being the head of the sentence. We give an example dependency tree in Figure 1.3. See also Figure 6.1 in Chapter 6.

Having the parse tree allows us to implement the linguistic noun-interpretation constraints provided by the field of Government & Binding Theory (outlined in Section 2.1). Furthermore, Minipar provides our anaphora resolution systems with noun-phrase identification, named-entity recognition, and some noun gender and number determination.

Syntactic structures not only provide information about which sentence elements govern other sentence elements, but they allow us to extract paths from the trees. Paths are syntactic segments with all modifiers and non-essential elements excluded. This allows us to reduce phrases like "John really needs his support" or "John needs his immediate support" or even "John will need his support" to the same core path between "John" and "his" depicted in Figure 1.3.

## 1.5 Outline of Thesis

In Chapter 2, we summarize related work in coreference and anaphora resolution, with a focus on both syntax-based and machine-learning approaches for reference resolution. We also look at other unlabelled corpus-based information extraction algorithms for reference resolution.

For our own approaches, a set of labelled data is used in the training of our supervised learners and testing of our algorithms. This manually labelled set provides an important resource for future coreference resolution work by the Natural Language Processing group at the University of Alberta. In Chapter 3, we describe this data and the unlabelled data used in our extraction procedures.

Two of our bootstrapped unsupervised extraction algorithms are described in the two subsequent chapters. In Chapter 4, we present our novel approach to extracting gender information by pattern matching in parsed corpora and pattern mining using an Internet search engine. This section also

describes the implementation of our multi-featured SVM anaphora resolution system. Chapter 5 describes our work in learning the likelihood of coreference along syntactic paths between potentially coreferent entities. These are extracted by looking for paths in parsed corpora linking pronouns of the same or different number/gender.

Any optimum bootstrapped extraction system needs to automatically handle pronouns that are non-anaphoric, which we define as those that do not corefer with a previously mentioned noun phrase. We describe these cases and motivate their handling in Chapter 6, which details our novel cataphora resolution algorithm, building on Chapter 5 by using antireflexive cataphora detection. We also present work in pleonastic pronoun detection.

In Chapter 7, we describe our fully-automatic, unsupervised Expectation Maximization approach to anaphora resolution. This work was done in collaboration with Computing Science doctoral student Colin Cherry. It uses the non-anaphoric pronoun handling modules explained in Chapter 6, and is compared to the supervised system of Chapter 4.

In our conclusion, Chapter 8, we summarize the main achievements of our work and state directions for future research.

# Chapter 2

# Related Work

For clarity, in each chapter we discuss research related to the specific approaches of that section. We also, however, provide here an overview of key research related to all of our methods. We begin with a discussion of anaphora resolution systems, including work in linguistics, early computational efforts, and later machine-learned approaches. We then discuss the notion of bootstrapping and its relevance to NLP and anaphora resolution in general.

For general summaries of anaphora resolution we refer the reader to the following thorough treatments: Mitkov [52] discusses the state-of-the-art in anaphora resolution at the end of the nineties. Jurafsky and Martin [32] summarize some of the important factors in anaphora resolution, and describe Hobb's algorithm, that of Lappin and Leass, and an approach called Centering Theory.

## 2.1 Government and Binding Theory

Although our primary concern is computational approaches, there is a strong linguistics framework for resolving anaphora. This framework has both influenced the direction of this research and provided syntactic coreference constraints for the two automated resolution systems outlined in this thesis (Chapters 4 and 7). The linguistics framework is called "Government and Binding Theory," and sometimes more comprehensively referred to as "Principles and Parameters Theory" [13, 23]. This work developed from Noam Chomsky's forty years of work in generative grammar.

Haegeman explains the Chomskian perspective on linguistics in her book [23]. In traditional language study, we might choose a specific language, for example English, and attempt to characterize all the principles that describe how sentences are formed – basically, what the grammar of the language is. Among other facts, we might observe that English sentences usually occur in subject-verb-object order. Linguistic research in the last forty years has sought not to develop this precise characterization of one language, but to develop the underlying principles that regulate the formation of all sentences in all languages. These principles are usually developed with respect to parse-tree representations of sentences – the syntax that forms the core of grammar. As mentioned in Chapter 1, computationally we are able to access this syntax using parsers such as Minipar.

Arguments for the existence of such universal linguistic principles are usually expressed with respect to biological or philosophical considerations. It is evident that each speaker of a given language must possess an unconscious internal representation of the syntax of their language, for they are able to both produce novel sentences conforming to their language's grammatical principles and decide whether given sentences are grammatical.[1] Furthermore, they achieve this proficiency at a young age, after being exposed to only a portion of the infinite variety of sentences possible in their native tongue. This is part of the so-called "poverty of the stimulus."

Linguists in the Chomskian tradition argue that a large part of a child's representation of their grammar must therefore be innate. And since any child is able to learn any language, this innate representation of grammar must be universally useful to all humans. Indeed, this is the justification for discovering underlying principles common to all languages. We are endowed genetically with portions of the brain reliably involved in language. Some of these innate brain functions are believed to encode principles of a so-called universal grammar. As a child learns a particular language, this universal grammar is configured to suit the language of exposure. Key groups of parameters are set at this stage of development, implicitly telling a speaker what kinds of sentences are acceptable through experience.

Many aspects of this theory remain in flux, including the extent to which the universal grammar or the learned lexicon of a particular language controls the syntax, but for our purposes, we can focus on one well-studied and important portion of the universal grammar that seems relevant to computational approaches in any language. This concerns the common mechanisms for the interpretation of nominal phrases in sentences. When resolving a pronoun, Government and Binding Theory shows how certain other noun phrases in the sentence cannot be coreferent. (Indeed, this prominent branch of the theory called Binding Theory has come to provide the name of the entire research program.) This is precisely the kind of knowledge we would like to incorporate into our systems to increase our computational accuracy.

Binding Theory gives constraints on coreference for reflexives, pronominals, and full noun phrases:

(1) John struck *him*. (*pronominal*)

(2) Mark said *he* struck *him*. (*pronominals*)

(3) John struck *himself*. (*reflexive*)

(4) He struck *Mark*. (*full noun phrase*)

Although a full exposition of binding theory is beyond the scope of this thesis, we provide a flavour of the theory as follows: Think of a pronoun's *governing category* (gc) as a small phrase

[1] Native speakers can say whether a sentence is grammatical even if it is non-sensical, as in Chomsky's famous declaration that "colourless green ideas sleep furiously."

containing both the pronoun and a subject noun. As well, note that a noun is *bound* by a preceding entity if this entity both corefers with the pronoun and commands it (command: its parent is an ancestor of the pronoun in the parse tree, but the noun itself is not an ancestor of the pronoun). The three principles of binding theory (using our own terminology) are:

- Principle A: A reflexive must be bound in its governing category.
- Principle B: A pronominal must be free in its governing category.
- Principle C: A full noun phrase must be free everywhere.

These three principles give us powerful tools when deciding what the potential candidates are for a given antecedent. In Sentence 1, Principle B tells us that "him" cannot refer to the noun phrase "John," since "John" is within the pronoun's governing category. On the other hand, coreference between any of the pronouns and "Mark" is possible in Sentence 2 because "Mark" is outside the gc. Here, in Sentence 2, the governing category of both pronouns is the phrase "he struck him," with the subject noun phrase being the pronoun "he." Anything outside this governing category can corefer with entities inside the governing category. On the contrary, reflexives *must* corefer with a commanding entity in their governing category, thus "John" *must* be the antecedent of "himself" in Sentence 3. Finally, "he" cannot refer to "Mark" in Sentence 4, since Principle C requires that full noun phrases must not be bound by anything.

Principle C applies less frequently in computational anaphora resolution than Principles A or B. Most systems look backward for candidate nouns, avoiding "Mark" automatically. However, "Mark" *might* occur in preceding sentences as well. Thus, our system rejects preceding nouns matching a Principle C-blocked forward noun phrase.

Determining the actual antecedents of the pronouns in Sentences 1, 2, and 4 is beyond the scope of Government and Binding Theory, as it is beyond the scope of grammar itself. Grammar only says what is possible within a sentence, other areas of linguistics, notably pragmatics and discourse theory, are interested in interpreting the utterances [23]. In essence, the approaches we describe in this thesis use statistics to determine what is most likely in a discourse.

There remain certain deficiencies in Government and Binding Theory relevant to our work. Based on the above principles, it is clear that reflexives and pronominals cannot occur in the same position and refer to the same entity. Yet sometimes this is in fact appropriate [27]:

(5) $\text{John}_i$ saw the picture of $\text{him}_i$ on the table.

(6) $\text{John}_i$ saw the picture of $\text{himself}_i$ on the table.

In these examples, no matter how exactly Binding Theory is defined and implemented as a constraint in our work, applying it to one or the other of the above cases would lead to an incorrect resolution. This remains an issue for computational implementations of Binding Theory.

Another computational issue that should be emphasized is the dependency of Government and Binding Theory constraints on correctly parsed sentences. In automatically-parsed resolution approaches like those described in this paper, your linguistic insight is only as good as your syntactic representation of the sentence. Fortunately, Minipar parses with relatively high accuracy. Future work should be conducted to quantify the accuracy of binding theory, both using manually parsed (correct) trees and those produced by Minipar. We discuss such work in our conclusion, Chapter 8.

In general, we have found the above constraints to be computationally quite consistent. As well, the fact they are applicable across a number of languages thus increases the scope of the tools we have developed using them. Chomsky [13] points out that "existing languages are a small and in part accidental sample of possible human languages." Furthermore, when searching for a universal grammar, it seems that any language may be as important as any other language insofar as its properties can provide clues to innate syntactic representations. Yet one might argue that it would be justified in computational linguistics to focus on a few prevalent languages to achieve top performance on the most important and widespread applications. One might also argue that taking such a narrow focus would reduce the scientific merit of the work. All else being equal, an approach that works across all human languages is much better than one designed for a specific form of communication. Although we designed our algorithms using English texts, simple refinements could enable every one of our approaches to be used in any other language with a written form and sufficient example text.

The field of linguistics provides more assistance to anaphora resolution than simply constraints on coreference. The ideas behind our pre-processing steps – the very syntactic structures we induce by parsing – have their foundation in linguistics. For example, our parser, Minipar, is based on both the Principles and Parameters theory of linguistics and also recent developments in Chomsky's Minimalist Programme [42]. Other coreference resolution systems also have their pre-processing foundation in linguistically-motivated parsers. As we turn to more computational approaches, it is worth keeping in mind the debt of each of these systems to the pure linguistics research that enabled their application.

## 2.2 Early Computational Work

We now briefly explain some previous (mostly) syntax-based approaches to anaphora resolution. Without some form of syntactic parsing, the resolution algorithm has no access to the linguistic constraints outlined above, and thus is suboptimal compared to the syntax-based approaches. Of course, there have been a number of suboptimal approaches published in the literature, simply because the effort or expense of parsing is not warranted for the given application. We mention some of these "knowledge-poor" approaches below, but mostly focus on systems leveraging a full parse-tree sentence representation.

In 1978, Jerry Hobbs developed a simple approach to resolving anaphora that relies on a tree-

search procedure on parsed sentences [27]. When searching for the antecedent of a particular pronoun, the Hobbs approach uses a left-to-right breadth-first search of the parse tree, stopping when it finds a noun that satisfies certain constraints. Hobbs called this the "naive" approach. In a manual test on 100 third-person pronouns (not including reflexives), assuming correct parsing, and removing pleonastic[2] cases, Hobbs achieved an accuracy of 88.3%.

There are two main issues with this evaluation: first, manual evaluation on special pronoun cases in correctly parsed sentences will reach levels of accuracy far beyond practical implementations. This is especially true of a syntax-driven approach. Secondly, the naive algorithm cannot handle the large number of cases requiring real-world knowledge to disambiguate between various antecedents. As Hobbs [27] explains:

> [These] results set a very high standard for any other approach to aim for. Yet there is every reason to pursue a semantically based approach. The naive algorithm does not work. Any one can think of examples where it fails. In these cases it not only fails; it gives no indication that it has failed and offers no help in finding the real antecedent.

Nevertheless, this approach has remained an important comparison algorithm for many proposed anaphora resolution methodologies. Lappin and Leass [38] compared their system to the Hobbs approach and showed 4% performance improvement on a data set comprised of technical manuals (86% versus 82%). Tetreault [71] compared the Hobbs algorithm to newer pronoun resolution methods and found it outperformed more complicated approaches. Ge et al. [21] used the so-called "Hobbs distance" – a measure representing the order the candidates are visited in the search – as a feature in their statistical resolution algorithm.

Aside from Hobbs, much of the early work on anaphora resolution was within the context of developing a general model of discourse structure and interpretation. By using properties of discourse, such as coherence and focusing, along with additional syntactic, semantic or pragmatic constraints, one can identify antecedents [38]. For example, Brennan et al. [9] developed a centering approach to pronoun resolution which is based on the model of attentional structure in discourse [22]. Basically, pronouns serve to focus attention on the topic of discourse. In fact, communication with pronouns removed or with pronouns used improperly is less fluent [9]. Given that pronouns are needed to provide topic continuity, keeping track of the topic (or center, or focus) of the dialogue in question provides a means to choose the antecedent for an ambiguous pronoun.

Mitkov [52] gives the following example: It would be difficult for anyone to resolve the pronoun in the phrase, "Jenny put the cup on the plate and broke it." However, if the preceding discourse was clearly focused on the cup, for example, then the true antecedent is clear:

(5) Jenny went window shopping yesterday and spotted a nice cup. She wanted to buy it, but she had no money with her. Nevertheless, she knew she would be shopping the following day, so

[2] Pleonastic pronouns are non-anaphoric instances of the pronoun *it*, such as in "*it* is raining." See Chapter 6 for further discussion.

she would be able to buy the cup then. The following day, she went to the shop and bought the coveted cup. However, once back home and in her kitchen, she put the cup on a plate and broke it...

In our own research, we incorporate topic focus by including the frequency of the candidate antecedent as a feature in our machine learned anaphora resolution approach (Chapter 4).

In 1994, Lappin & Leass [38] presented an influential algorithm for determining the antecedents of pronouns called the "Resolution of Anaphora" (RAP) algorithm. RAP scores various features such as recency, frequency, subject emphasis, parallelism, etc., and chooses the highest scoring candidate from a list of antecedents. It contains many of the key components needed for a fully-automatic approach, including parsing and various syntactic and morphological filters. Syntactic filters are based on McCord's Slot Grammar, and prevent coreference in a manner similar to our linguistic constraints above. Morphological filters look at features of the noun and pronoun, and prevent coreference for entities of different person, number or gender. Their parameters were optimized for the domain of computer manuals, and they achieved 86% accuracy on 360 pronouns in a blind test. A parser-free re-implementation of this approach by Kennedy and Boguraev [35] achieved 75% performance on 306 anaphoric pronouns on a variety of texts including news articles. Both the Hobbs algorithm and the Lappin & Leass algorithm are discussed in [32].

By the late 1990s, it was standard practice to subject coreference and anaphora resolution work to empirical evaluation. Most of this work was not based on learning from a reference-annotated corpus ([3, 33, 38, 35]). This was also the era when the Message Understanding Conferences [55, 56] had established coreference resolution as an important and distinct problem from anaphora resolution (as mentioned above, concerned with partitioning all noun phrases into coreferent groups, beyond only resolving anaphoric pronouns). However, resolving pronouns remained an important part of the coreference resolution problem. Lin [40] used linguistically-motivated features such as Binding Theory to resolve pronouns in his PIE system for MUC-6.

Development of so-called "knowledge-poor" approaches also became popular in the late 1990s as a way to avoid the time-consuming and labour-intensive acquisition of linguistic and domain-specific knowledge [52]. One common characteristic of these systems is the absence of a full syntactic parser. Full parsing of text may be especially problematic in information extraction applications, such as searching on the world wide web, where parsing all the documents in cyberspace before reference resolution may be impractical. Thus, the goal has been to find ways to do robust reference resolution with "impoverished syntactic analysis" of the input text [33].

The aforementioned approach of Kennedy and Boguraev [35] falls into this category. Breck Baldwin's CogNIAC system [3] is also considered knowledge-poor. It works with part-of-speech tags, and is notable because it leaves pronouns lacking a clear antecedent unresolved. The term "knowledge-poor," however, is most often associated with a series of systems by Mitkov and others, culminating in a fully-automatic, knowledge-poor approach detailed in [53]. This system reaches

62% on 2263 anaphoric pronouns (excluding pleonastic pronouns – results are also stated for the automatic pleonastic-handling case).

See [52] for more examples of knowledge-poor or syntax-free systems.

### 2.2.1 Evaluation Issues in Anaphora Resolution

Careful readers will have noticed the seemingly baffling trend of poorer and poorer performance results in resolution research as time progresses. This does not reflect any inherent degradation in the design of algorithms, but rather a move from purely simulated to purely automatic systems.

The majority of previous anaphora resolution approaches involve some form of manual involvement [53]. The approach of Hobbs [27] was not implemented but manually simulated. Ge et al. used a manually-parsed corpus [21], while Lappin and Leass manually corrected parser output [38]. In all of our work, we parse the text and perform noun-phrase identification fully automatically.

Lappin and Leass had a module for automatic identification of pleonastic pronouns [38], while Kennedy and Boguraev manually identified and excluded these pronouns, as well as those that refer to verb phrases or sentence clauses [35]. In Chapter 4 we also remove these pronouns automatically, while Chapter 6 presents a method to automatically handle these pronouns, which was essential for the unsupervised approach of Chapter 7.

Different levels of manual involvement and manual versus automatic handling of special pronoun cases is partly what makes comparison of anaphora resolution approaches so difficult. Despite some efforts at standardization, use of different data sets and different levels of preprocessing is common practice in anaphora resolution applications.

However, there are some advantages to a lack of standardization. Different researchers working from different positions encounter different phenomena. This contributes significantly to the overall development of the field – rather than endlessly tuning and re-configuring systems to score higher on the "standard" sets, as has happened in other NLP areas, researchers learn more about the wide range of issues relevant to anaphora that cannot be encompassed in a single labelled corpus.

Lack of a common testing ground is not an important issue for most of the research covered in this thesis. Our overall anaphora resolution methodology is not primarily intended to compete with other systems, but rather to show the relative benefit of incorporating new constraints and probabilistic information. That being said, we have ourselves labelled a large amount of text, and invite other researchers to both use this work for their own development and to compare with our approaches (see Chapter 3 for details).

Also note that "baseline" numbers, such as the performance of always selecting the previous noun phrase, provides a fairly consistent way to compare the inherent difficulty of the text used. Although use of this baseline statistic is not widespread in the field, we encourage its use and state results using it on all of our data sets. Although not without its own drawbacks, a *relative improvement over previous noun heuristic* might provide a consistent way to compare the performance of

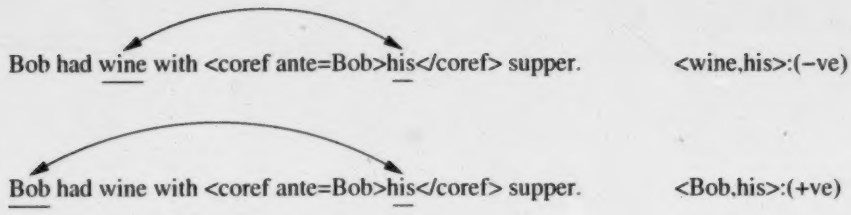



Figure 2.1: Example feature vector creation.

different algorithms on different text data.

## 2.3 Machine Learning Approaches

Although empirically-motivated, many of the early approaches to anaphora resolution required manual configuration of parameters, and thus their applicability to new domains, styles, and languages was limited. However, a clear scoring function (resolution accuracy) was available to evaluate the algorithms upon their application to new data. This would have allowed for easy testing of new sets of parameters as the algorithms were tried in new areas. It perhaps now seems obvious to employ machine learning to re-configure these systems when the application area changes. Yet learning from an annotated set of coreference relationships was not adopted before 1995. We discuss such approaches in this section.

### 2.3.1 Machine Learning for Anaphora Resolution

Supervised machine learning requires a set of feature vectors and a set of class labels. The learning module decides how to incorporate the given features into a function that distinguishes between the given classes. In anaphora resolution, each of the feature vectors represent a possible coreference relationship. The learner decides how to combine the features of a possible antecedent-anaphor pair to distinguish between coreferent and non-coreferent entities (i.e. a binary classification task).

The features of a possible antecedent-anaphor pair can be any of the standard features used for anaphora resolution, e.g. gender/case/number match, whether the antecedent is a subject or not, whether the syntactic parents of the pronoun and antecedent match, etc. Once a feature representation is chosen, positive and negative instances (coreferent or non-coreferent pairs) must be extracted from the annotated data to provide the training set for the classifier.

To create the training set, the most sensible procedure we have observed, and the one adopted in our work, is the procedure of Soon et al. ([69]). Each pronoun and its closest preceding antecedent in our labelled set of training documents form a pairwise positive instance in the set of training vectors. All intervening noun phrases (between the antecedent and the pronoun) form pairwise negative instances with the pronoun. Figure 2.1 illustrates the process for creating two feature vectors for a given labelled pronoun.

The feature vectors created in this manner are used to learn a classifier, which is then applied incrementally backward from each anaphor in the test set until a given antecedent is accepted as coreferent with the anaphor. Accuracy is defined as the percent of antecedents correctly identified in this manner. We shall discuss other evaluation metrics in future chapters.

The process sounds simple enough and fairly efficient. Why configure parameters manually when they can be learned from annotated data? Yet we shall see that there are only a handful of such approaches in the literature.

### 2.3.2 Previous Approaches

Before the turn of the millennium, the only works in learning of coreference or anaphora resolution from an annotated corpus were by Aone and Bennett [2], Fisher et al. [20], McCarthy and Lehnert [47], McCarthy [48], Kehler [34], and Ge, Hale, and Charniak [21] (according to Soon et al. [69]). Of these, only Ge et al.'s approach focused solely on anaphora resolution.

#### Coreference Resolution Approaches

The RESOLVE system is described in [20, 47] and [48]. It was one of the first to use decision tree learning. Soon et al. [69] report that the disadvantage of this early effort was the domain-specificity of the feature set (focused on identifying coreference in the domain of "joint ventures"). Machine learning can tune a system for a new domain only so far as the feature representation can capture the criteria for coreference in that new domain.

Aone and Bennett [2] also used decision trees. They worked on Japanese texts, and only evaluated nouns referring to organizations. Also, they used noun phrases that had been correctly identified. Again note that manual involvement was a hallmark of most early approaches to anaphora and coreference resolution.

Kehler used Maximum Entropy to assign a probability distribution to possible coreference relationships [34]. In Chapter 7 we also use Maximum Entropy to re-weight probability models used to choose the antecedent from the list of candidates. And, like Kehler, our approach returns a probability distribution over possible candidates, rather than a hard decision on a coreference link. Probabilistic coreference information, rather than hard decisions, may be more useful to downstream applications relying on the coreference information.

Cardie and Wagstaff [11] also had a machine learning approach, but it was based on unsupervised learning via clustering, and thus did not require annotated data. They achieved predictably inferior results to supervised learning; however, whether the effort in annotation justifies the gain in performance will naturally depend on the application in question.

Soon et al. [69] were the first to show competitive results with machine learned models compared with state-of-the-art non-learned systems. They used a fully-automatic system, with domain-independent features. Supervised decision tree learning was used to induce the coreference res-

olution classifier. They achieved truly impressive performance given the limited number of features available to their classifier. The following year, Ng and Cardie [59] added a larger set of linguistically-motivated features to the Soon et al. baseline system, and were able to achieve superior performance. However, they had to "hand-select" a subset of these features in order to prevent overfitting with the decision tree learning algorithm.

**Anaphora Resolution Approaches**

Coreference resolution systems handle pronouns as a subset of their noun phrase coverage, but generally only devote a limited number of features to pronoun resolution, and achieve results inferior to those of the manually-assisted systems devoted solely to pronouns. It seems likely that superior performance could be achieved on the pronoun subset by learning separate classifiers with features more relevant to pronouns, and training on the pronouns individually.

Ge, Hale and Charniak [21] learned a statistical model of anaphora resolution from labelled text, and used it to resolve pronouns only. They reached very high levels of performance with this approach, up to 84.2%, although it relied on a manually-parsed data set. They also used this system to label a large amount of unlabelled text, bootstrapping gender knowledge in a manner similar to our own approach. We discuss this aspect of their work in the following section.

One of the first approaches to mine anaphora resolution information from corpora is by Dagan and Itai [15]. They looked at certain co-occurrences in text to determine the suitability of each of the candidate nouns given the pronoun's context. For example, the verb parent of a pronoun may be used to select antecedents that satisfy the verb's selectional restrictions. If the verb phrase was "shattered it," we would expect "it" to be coreferent with some kind of brittle entity, and we could filter the candidate nouns accordingly. By looking for the nouns that occur as objects of "shatter" in text, Dagan and Itai's approach is able to automatically discover what entities might be considered brittle. Similar information was incorporated in [38] and [21] where leveraging such statistical data was only shown to provide roughly a 2% performance improvement. Ge et al. [21] suggest a number of possible reasons for such a small difference; ultimately it depends on the genre of text being resolved. In news texts, for example, the verb "say" is very frequently the parent of the pronoun and it has many possible subjects. Knowing which entities can "say" things provides little help beyond the information already encoded in the person/number/case of the pronoun and antecedents.

## 2.4 Bootstrapping

In general, bootstrapping is a way of taking steps toward a solution, then using the results of those steps to achieve your final goal. In the past, a bootstrap was a strap attached to the top of the boot that helped you pull your boot onto your foot. The step of pulling on the strap facilitates the ultimate goal of pushing your foot into the boot. Dave Winer [74] explains bootstrapping with the following example: when engineers build a bridge, first they draw a single thin cable across the water. They

then use this cable to support the weight of larger ones that are subsequently drawn across. Then these larger cables are used to pull even larger ones, which eventually are strong enough to support the full weight of the men and machinery needed to build the bridge. Note that when you "boot" your computer, you load the first piece of software into memory necessary to call and run all the other programs your system depends on.

In Natural Language Processing, bootstrapping learning has often meant using a portion of labelled data to bootstrap learning on a large amount of unlabelled data. Co-training [8] works in this fashion. Two separate views of the data to be classified are employed, for example, two different feature vector representations. We might represent a web page by the words on the page, or by the hyperlinks pointing to the page. We can learn a classifier for each view, and then use these classifiers to generate more training data for the other view. Blum and Mitchell [8] gave a PAC Learning-style framework for this approach, and gave empirical results on the web-page classification task.

Initially, co-training was not found to be successful for anaphora resolution, where no natural view factorization is available [57]. Ng and Cardie [61, 60] showed greater success with single-view co-training algorithms, which instead of separate views, use two separate learning algorithms.

In our view, a bootstrapping process for anaphora resolution does not necessarily require labelled data to bootstrap the resolution process. We perform bootstrapping by performing resolutions and learning from the aggregate data collected from the total resolution dataset. This information is fed back into an anaphora resolution system.

In this sense, one might also term the approach of Ge, Hale and Charniak [21] bootstrapping, since earlier resolutions are used to extract gender and number information, which in turn is fed back into the coreference resolution system. Similarly, Bean and Riloff [5] used extraction patterns based on assumed coreference links to gather contextual role knowledge for anaphora resolution. Contextual role knowledge determines whether the contexts around an anaphora and potential antecedent are compatible. For example, if one sentence describes kidnappers and their victims, the antecedent of the pronoun "they" will be different in the phrase "they were released" than it would be in the phrase "they blindfolded the men." This is similar to the context information mined by Dagan and Itai [15] and the knowledge we incorporate as mutual information in Chapter 4 and as a language model in Chapter 7.

Harabagiu et al. [24] also learned coreference constraints from unlabelled text using a bootstrapping approach. They applied a coreference resolution system learned from an annotated set to a large corpus of unseen text. The semantic consistency of coreference relationships created in the new data was inspected to determine novel rules for common noun coreference. Although the aims and method of their work were different, the spirit of assuming or creating coreference links in text automatically and using these unvalidated decisions to deduce general knowledge is the essence of bootstrapping.

# Chapter 3

# Data Sets

In this chapter we explain both the annotated data used for supervised learning and testing, as well as the large repositories of unlabelled text used in our unsupervised extraction algorithms.

## 3.1 Annotated Data

One of the main contributions of this work has been the annotation of anaphoric relationships in a large amount of text data. This data will provide a test set for any future anaphora resolution work in the Natural Language Processing group, and remains useful training data for supervised approaches to anaphora resolution.

We labelled third-person pronoun-antecedent pairs in 118 documents from the slate section of the American National Corpus (ANC) and 176 documents from the New York Times, Associated Press and Xinhua news portions of the AQUAINT [sic] corpus. In the present thesis, the annotated ANC data provides the training data and separate testing data used by both the gender classifier described in Section 4.5, and the full anaphora resolution system described in Section 4.6. This training data is also used to create the comparison SVM system used in Section 7.7.1. The annotated AQUAINT data is used to create the development and test keys used in our unsupervised approach to pronoun resolution (Section 7.6).

### 3.1.1 Annotation Notes

We use XML-style tags to label the antecedent (or lack thereof) for all pronouns in our annotated set. For example:

(1) The law defines political activity to include anything campaign-related –organizing events, planning party strategy– except fund raising , which <coref ante="law">it</coref> completely prohibits.

We labelled antecedents without knowing exactly how previous nouns will be parsed by Minipar. We thus provided fairly descriptive labels in each tag – usually all the nouns in a given noun phrase,

that is, without any determiners, prepositions, or other non-nominal modifiers included. When deciding whether a resolution has occurred, we accept preceding nouns that match the final delimited portion of the label, provided this portion is separated by a space character, and also provided neither is a conjunction. For example, we would declare a match if the antecedent was "Clinton" and our label was "Bill Clinton," but no match for the antecedent "Clinton" if our label was "Gore and Clinton." Furthermore, "port" will not match "airport" but "port" and "air port," with a space delimiter, will be matches. After running our various anaphora resolution programs several times on all the annotated data, we fixed a number of labels to properly match and prevented some that were matching incorrectly. However, sometimes the preceding antecedent is simply not parsed as a noun, or not parsed correctly (i.e. the head of the noun phrase is not included in the parse, but rather parsed as a verb or other entity). In these cases, our anaphora resolution algorithm may still find an earlier occurrence of this entity to match with, but in a few instances, no previous noun is found to match, and our system scores unavoidable errors.

As we shall explain in detail in Chapter 6, not every pronoun in text is anaphoric. We label pleonastic pronouns with the "NULL" symbol as antecedent, and non-noun referential pronouns with an "IMPLICIT" label:

(2) And what would <coref ante="NULL">it</coref> cost to make the promise credible?

(3) On the sixth day, God created man in <coref ante="god">His</coref> own image: Adam was produced from dust, and Eve from Adam's rib.
God told <coref ante="IMPLICIT">them</coref> to reproduce and rule over the world's living things.

Chapter 6 also deals with the non-anaphoric pronouns called *cataphora*. Cataphoric pronouns are those occurring *before* their antecedents. We also label genuine cataphora in the ANC section of our labelled set, but not in the AQUAINT section. This is because no cataphora-handling module was in place at the time the ANC data was to be used. With cataphora labelled in the ANC text, however, we were able to inspect a number of instances of cataphora automatically, which assisted when developing the cataphora-handling algorithm outlined in Chapter 6 and used in Chapter 7.

Of the 2779 total pronouns labelled in the ANC data, 144 are pleonastic, 59 do not refer to an explicit noun phrase, and 16 are cataphora.

### 3.1.2 ANC Data

The first release of the American National Corpus was made available in the fall of 2003 and contains about 11 million words of American English [28]. Ultimately, the ANC aims to serve as a resource for language and linguistic research similar to the British National Corpus (BNC). Details on obtaining the corpus are available at its website (http://americannationalcorpus.org/).

We chose to annotate *gist* articles from the *slate* section of the American National Corpus. These articles provide factual background information for stories currently in the news. Such informative articles are an excellent source of knowledge for question-answering applications, ensuring our research efforts are focused on resolving anaphora with this ultimate application in mind. Many previous approaches did not tackle news texts, but instead, for example, computer manuals and other more structured genres.

Our labelled ANC documents are divided into two sets – one for testing and one for training. There are 1398 labelled pronouns in 79 documents in the training set and 1381 labelled pronouns in 41 documents in the test set (including non-anaphoric cases).

Unfortunately, the first release of the ANC contains some systematic errors in the *slate* section of the corpus – including incorrect sentence breaks and missing letters at the beginning of certain paragraph-initial sentences. We add tags to mark where we have fixed the systematically missing letters. We also add tags around bad sentence breaks. The sentence break tags allow us to re-merge the two portions, and our LaTaT tools can then use their own sentence boundary detection to divide the sentences. These LaTaT sentence boundary detectors operate with a much greater efficiency than the tools used to partition sentences in the original articles, although occasionally they make the same errors, which we accept. Note, we have been told that the second release of the ANC will correct many of the errors visible in the first version.

We have sent all of our ANC annotations to Nancy Ide and Keith Suderman at the ANC, so that they may be shared with the wider NLP community. Keith has extracted out the anaphora reference labels and provided "stand-off annotations" for all of the documents we marked up. These will be available with the second release of the corpus. Tools have been developed so that anyone with access to the second release will be able to merge the annotations back into the text automatically. For those with the first release, it is still possible to merge the annotations in their current form with the *slate* articles we used. Some basic instructions for obtaining and using the stand-off annotations are available at:

http://www.cs.ualberta.ca/~bergsma/CorefTags/.

### 3.1.3 AQUAINT Data

The AQUAINT Corpus is the corpus used in the evaluation portion of the TREC Question Answering track [73]. Its full name is the AQUAINT Corpus of English News Text. It is available from the Linguistic Data Consortium (www.ldc.upenn.edu). It consists of approximately one million documents taken from the AP newswire from 1998-2000, the New York Times newswire from 1998-2000, and the English portion of the Xinhua News Agency newswire from 1996-2000. The full corpus contains roughly 3 gigabytes of text.

Because using anaphora resolution to improve question answering remains an exciting research area, we felt the anaphoric-annotation of portions of AQUAINT may have many future applications

for ourselves and other researchers. We again annotated both a training portion and a test portion (later referred to as the "development key" and "test key" in connection with their use in our unsupervised evaluation, Chapter 7), both consisting of numerous documents from each of the three news agencies. The training set consists of 644 labelled pronouns drawn from 58 documents. The test set consists of 1209 labelled pronouns drawn from 118 documents.

## 3.2 Unlabelled Data

Although the labelled data is essential for testing our approaches and training our supervised machine learning algorithms, our methods of leveraging unlabelled data remain the most important contributions of this thesis. Most of our bootstrapping approaches result in information whose value is directly proportional to the quantity of text used in the learning process. Thus, a large amount of text is incorporated.

In the approaches outlined below, the main sources of data are:

- The full American National Corpus [28].
- The full AQUAINT corpus [73]
- the Reuters corpus [66]
- The Associated Press, Wall Street Journal and San Jose Mercury News sections of the TIPSTER corpus [26]

Together, the full set contains about 6 gigabytes of text. Again, note the domain of extraction is largely newspaper articles, ensuring the collected data will be applicable to the newspaper articles on which we test our resolution algorithms.

# Chapter 4

# Extracting Gender and Number Information

## 4.1 Introduction

In this chapter, we explore an approach for determining the antecedents of pronouns using enhanced statistical gender and number information, based on our work in [7]. The key idea is that very reliable probabilistic noun gender information can be extracted automatically from text. We show significant gains in performance by using this probabilistic data within baseline and machine learned anaphora resolution strategies.

The basic flow of our algorithm is depicted in Figure 4.1. We gather gender using an extraction process outlined below, and store the information for each noun lexically in a database. This probabilistic gender lexicon can then be used as an input to assist in classifying anaphora, as we see later in Figures 4.2 and 4.3.

Why is gender/number information useful for anaphora resolution? As we explain below, each pronoun in text has an explicit gender/number – one of masculine (e.g. *his*), feminine (e.g. *her*), neutral (e.g. *its*) or plural (e.g. *their*). We alternately refer to these four kinds as "gender/number," "noun category" (Chapter 7), or simply as "gender." Having four distinct noun categories into one of which each referenced entity must fall is closer to the linguistic sense of "gender," meaning grammatical gender or noun class (of which some languages have many more than the four we use here), rather than the sense of gender as "sex," a more recent usage of the term [65].

When we are looking for the antecedent of a particular pronoun, we can use the gender and number information to eliminate candidate nouns from consideration as antecedents. Section 4.2 further explains the use of gender and number match as an anaphora resolution constraint and summarizes related gender determination strategies.

Section 4.3 describes the parsed-corpus and web-based gender-gathering templates which we use to extract the probabilistic gender information, and how the probability of a given gender can be modelled from these templates using a *Beta* distribution.

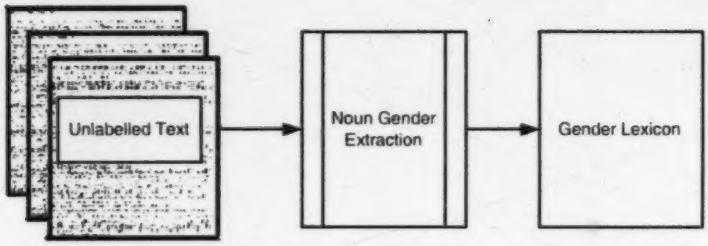



Figure 4.1: Gender information extraction process.

Details of our experimental set-up are given in Section 4.4. We evaluate our gender information through both pure gender-determination exercises (Section 4.5) and as features in full anaphora resolution systems (Section 4.6).

## 4.2 Gender in Anaphora Resolution

In many languages, including English, pronouns reflect the natural or grammatical gender/number of the entity they reference. Thus when determining the antecedent of a particular pronoun, it helps to know which of the preceding nouns are compatible in their gender/number class. Indeed, gender and number agreement are currently standard constraints in virtually all competitive anaphora resolution systems. As humans, we often require gender information to disambiguate between possible referents for an anaphor. Consider the following examples:

(1) Mary never saw Fred at the airport. He was on a different flight.

(2) Mary never saw Fred at the airport. She was on a different flight.

In both instances, we interpret the sentences differently solely based on the difference in gender between the two pronouns. Note, however, that to make these resolutions, we need to have an intuitive sense of the inherent gender of the words "Mary" (feminine) and "Fred" (masculine). This is the information we would like our system to automatically extract from unlabelled text.

There are a number of strategies for determining the inherent gender of a given noun. If we notice in text that the noun has a gendered designator, such as the noun "Mr. Bean," we can immediately assign a particular gender to this word. Pronouns, of course, also reflect gender, and gender constraints can be applied to eliminate from antecedent consideration pronouns coming from non-compatible categories. Some words in English have suffixes that can indicate gender, such as "seamstress" or "chairman," but the reliability of these suffixes is decreasing with time [19].

Our parser, Minipar, can determine gender/number information for some words. Minipar contains a list of gendered first-names, which it uses to assign male/female gender to unambiguous gendered-name nouns. Minipar, like many other parsers and part-of-speech taggers, is also able

to determine when a noun is plural based on the morphology of the string. Parsers leverage this information to improve their parsing accuracy, but the leveraging of gender information is not as widespread [19].

The above strategies might be called the traditional or "standard" approaches to using gender in anaphora resolution. Gender information is only used when it is clearly available from the text. As well, as we have been discussing, this information is used as a filter on candidate antecedents. Where our approach differs from most previous work is that we assume a word does not have one true gender, but a gender distribution over the four gender categories. When this gender distribution is determined, it can be used probabilistically within anaphora resolution systems.

For example, words like "president," "nurse," or "farmer" can be referenced by both male and female pronouns, but some of these references are more likely than others. We seek to determine the likelihood that each of these words and all other nouns can be referenced by a masculine, feminine, neutral or plural pronoun. This is done by looking at the frequency of association between nouns and gendered pronouns in gender-indicating patterns in text.

This approach is similar in spirit to Ge et al. [21], who first attempted to extract gender information from unlabelled text. Their methodology, a form of bootstrapping, was discussed in Chapter 2. First, a simple pronoun resolution system, working without gender information, is applied to a large amount of unlabelled text. A noun is assigned the gender of whichever pronoun is most often resolved to it. This assignment is tested on a list of proper names with gendered designators, and is found to be correct in about 70% of cases.

We also mention one other gender-determination strategy gaining in popularity in the coreference resolution community – the use of WordNet [50]. By looking at the senses of a noun in WordNet, and the relation of this noun to other nouns, it is possible to constrain gender in certain cases. For example, if an entity is deemed to be a kind of object, then we can restrict reference to this entity to be from a neutral pronoun. The assumption is that *singular* objects would never be referenced by "he," "her," "they," etc.

Unfortunately, WordNet includes many infrequent senses of words which are quite unhelpful to anaphora resolution classifiers. For example, the nouns *dog*, *computer* and *company* each have a WordNet sense that is a hyponym of *person* [63]. Thus gender-match constraints relying on WordNet might allow each of these as the antecedent of the pronouns "he" or "she," not distinguishing them from other, possibly extremely more likely antecedents. Researchers can partly deal with this problem by only abstracting gender constraint information from the most frequent senses in the WordNet synset.

## 4.3 Gender Information Sources

### 4.3.1 Pattern Matching for Noun-Pronoun Gender Pairs

In [7], we describe how both parsed-corpus and web-mined patterns can be used to extract gender information. The key ideas are as follows:

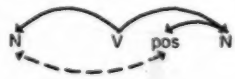
We saw in Chapter 2 how Principle A of Government and Binding Theory immediately gives the antecedent of a pronoun in phrases with reflexives. For example, for the phrase "the president introduced himself," we know that "president" and "himself" are coreferent in this context. Thus we can count this as one instance of "president" being masculine, since the referring pronoun is masculine. Each noun can be assigned a gender probability based on the frequency of coreference in this pattern with pronouns of a given gender class.

We have determined several other patterns that reliably link a bound noun and pronoun. The full list includes constructions with reflexives, possessives, nominatives, predicates and designators. We depict these below:

1. Reflexives (*himself, herself, itself, themselves*):

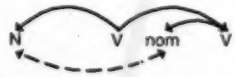
E.g. *Mary* explained *herself*...

2. Possessives (*his, her, its, their*):




E.g. *IBM* bought *its* supplies...

3. Nominatives in *finite* sub-clauses (*he, she, it, they*):




E.g. *Alice* thought *she* should...

4. Predicates: pronouns are subjects and nouns are in the predicate position:

E.g. *He* is a *father*.

5. Designators: The noun is accompanied by a gendered designator:

E.g. *Mr. Brown*.

Our first approach is to extract these patterns by parsing unlabelled text with Minipar. Everytime we encounter one of these pattern, we store gender information for the noun component based on the referencing pronoun in the pattern. Obviously extracting via this pattern is not always helpful. "Mary" may like "his" style – incorrectly giving the gender of "Mary" as masculine. Sometimes an error can occur because of a bad parse – due to the limitations of the parser or just because of ungrammatical text. Despite these concerns, given enough instances, a large enough corpus from

which to mine, and a random level of noise, the most likely gender for a given noun will usually have the highest number of pattern counts.

Given that we might benefit from a large amount of text, using the web to search for these patterns arises as an intriguing possibility. Rather than mining and extracting the patterns beforehand, we might search for relevant patterns on the web when we seek the gender of a new noun, and use the returned number of pages as an indication of the prevalence of that noun occurring as that particular gender. That is, the counts now do not refer to the number of times the pattern was extracted from unlabelled text, but the number of pages returned for that pattern on the web, where the four different pronoun genders are used with the five different constructions.

Like the corpus searches, the web-mined patterns will be noisy. Although we used the wildcard operator as a substitute for a verb, it is not restricted to be so. Also, the entire query string is not restricted to be in the same sentence. Finally, because our pairings tend to identify nouns in subject positions, we obtain limited data for nouns preferring object positions. Despite these limitations, we show the web-mined data outperforms the corpus-based information 4.5.

### 4.3.2 Combining Gender Information

At this stage we have gender information from five parsed-corpus sources and five web-mined sources. A noun can be given a gender probability based on any one of these sources. This section addresses how a single determination of gender can be made from this information array.

Individually, the sources may be noisy. For small counts, they may not closely approximate the true gender of the given noun, and in some instances, no patterns (or web-pages) may be found for that noun at all. We would like a way to both smooth the data when little or no counts have been obtained, and to obtain a measure of "confidence" in the counts we have received (a simple measure would be linear: the more counts we have, the more confident we should be). This confidence value can then be provided as a feature to our gender and anaphora resolution classifiers, which these classifiers can use to estimate how much confidence to place in the probabilities from a particular source. In this way, the classifier can dynamically rely more on the sources having high confidence when presented with the confidence and probability values for a particular noun. To provide the smoothing and to gather the confidence information for our sources, we adopt a Bayesian approach to modelling the individual sources.

In Bayesian parameter learning, a hypothesis prior distribution is assumed for the parameter, and this distribution is updated as new information is available [67]. In our work, the parameter is the probability this noun is of a particular gender. We initially assume any value for the gender probability is equally likely, in this way initializing the *Beta* distribution to a uniform prior distribution over the four genders. Subsequently, we treat this prior distribution as the first prior in a family of *Beta* distributions. Each count we find of a particular gendered pattern (or instance of that pattern on a web page) will sway the distribution probability toward that gender.

Generally, *Beta* distributions are used to model binomial proportions. For a given gender, we treat the pair counts as binomial in that all pairs of that gender are treated as one event, while any pair not of that gender is considered a separate event. The mean of the *Beta* distribution corresponds to a smoothed estimation of the probability that noun will occur as that gender. The variance of the *Beta* distribution provides the confidence measure. For *Beta* distributions, the more counts, the smaller the variance. Interested readers are referred to [7] for further details.

Although this may sound complicated, we should emphasize that what we get from the *Beta* distributions is exactly what we would have gotten had we gone with the maximum likelihood value for the gender, based purely on the count statistics, but incorporating add-one smoothing and the confidence measure from the variance. In our testing, neglecting smoothing or the confidence measure results in small but consistent decreases in performance.

Again, using the above approach, we are able to get a probability value and confidence measure for each of the five parsed-corpus and five web-mined sources. The question remains, how do we combine these sources together when estimating gender?

We do this by treating the means and variances as dimensions in a feature space, and we use machine learning to induce a classifier over a training set of these values and their corresponding genders. Each of the ten sources will provide a mean and variance value, yielding a 20-dimensional feature space.

Consider the case of determining whether a particular noun is masculine or not. A machine-learned masculine classifier can inspect the features corresponding to this noun, and decide whether the noun is acceptable as this class or not. A low mean or variance for a particular source would correspond to a low probability of being masculine, or low confidence in the probabilities, respectively, for that noun.

We look at gender-guessing tests, with machine-learned classifiers assessing noun gender based solely on the gender statistic features, in Section 4.5. In Section 4.6, we use the means and variances along with other features in a full machine-learned anaphora resolution system.

## 4.4 Experimental Details

For the acquisition of the gender information, we collect noun-pronoun patterns from the full set of unlabelled data outlined in Chapter 3, excluding the portions of the ANC corpus used in our experiments. We extracted over 4 million reflexive pairs, 32 million possessives, 28 million nominatives, 5 million predicates, and 17 million words with gendered designators.

For the training data and separate testing data used by both the gender classifier described in Section 4.5 and the full anaphora resolution system described in Section 4.6, we use the annotated ANC data described in Section 3.1.

A key realization was that the labelled anaphora can also provide a set of gender-labelled nouns. For each pronoun labelled with a particular noun antecedent, we know the gender of the noun in

Table 4.1: Noun Gender Classification Performance (%)

| | Precision | Recall | F-Score |
|---|---|---|---|
| masculine | 88.2 | 95.2 | 91.6 |
| feminine | 98.2 | 70.9 | 32.4 |
| neutral | 93.0 | 93.7 | 93.3 |
| plural | 98.6 | 89.8 | 94.0 |
| micro-avg. | 93.9 | 90.6 | 92.2 |

that sentence – it must match the gender of the referring pronoun. We extract all pronouns with nominal antecedents from the training and test sets, getting lists of 903 nouns and their genders and 876 nouns and their genders, respectively. About 24% of the nouns in the lists are masculine, 7.6% feminine, 33.8% neutral and 34.6% plural.

Recall that the majority of anaphora resolution approaches involve some form of manual involvement [53]. In this chapter, we use the ANC data, described in Chapter 3. It has manually-identified pleonastic pronouns, as well as manually-identified pronouns that do not refer to preceding noun phrases, including cataphoric pronouns. Recall that of the 2779 total pronouns labelled, 144 are pleonastic, 59 do not refer to an explicit noun phrase, and 16 are cataphora. Results stated below do not include these excluded cases.

For all machine learning tasks, we use the Support Vector Machine (SVM) learning algorithm with the SVM$^{light}$ [31] implementation. A linear kernel is used. SVMs are known to achieve very high performance on a range of learning tasks [31]. They have been used previously for Japanese pronoun resolution [29]; we believe this to be the first use of SVM with English anaphora resolution.

## 4.5 Noun Gender Classification

Given the list of nouns and their genders derived from annotated training data, plus our corpus-based and web-mined gender information, we can now learn and test pure gender classifiers. Note at this stage we are not resolving the anaphora – our only goal is to see how well our probabilistic information can be used to guess the gender of the nouns in the gender list. The resulting performance will give us an indication of how useful our statistical gender information might be in the overall anaphora resolution system.

We formulate our gender classification as a binary task: we learn separate classifiers for masculine, feminine, neutral and plural nouns. Learning is performed on the gendered nouns from the training set. Classification is performed on the gendered nouns in the test set. The particular classifier is presented with a noun, and must determine whether that noun can be referenced by a pronoun of the classifier's gender. Recall, each classifier works with twenty features – ten corpus based mean and variance values, plus ten web-mined mean and variance values. We use precision, recall, and F-score to assess the quality of the classifiers. Overall performance results are given in Table 4.1.

Testing on this set of gendered nouns is problematic in the sense that nouns can have more

Table 4.2: Miscellaneous Micro-averaged Noun Gender Classification Performance (%)

| | Precision | Recall | F-Score |
|---|---|---|---|
| Parsed-Corpus Features | 90.9 | 80.6 | 85.4 |
| Web-Mined Features | 92.4 | 88.6 | 90.4 |
| Average Human Performance | 88.8 | 88.8 | 88.8 |

than one gender, and the nouns in the list are not guaranteed to represent the most likely gender. Thus even if our classifier has perfect information, it will not be able to score correctly on every noun in the list. Another way of stating this is to observe that the labelled data set has given us gender for the particular noun *token* rather than for the noun's overall *type*. Yet when we perform anaphora resolution, we will also be working with noun tokens rather than noun types. Therefore our performance on this task will be indicative of our potential proficiency at the overall problem.

In the overall system (and in the systems mentioned below), the female classifier consistently achieves much lower recall, while the male classifier consistently has slightly lower precision. These two phenomena are not unrelated. Many nouns can at times be both masculine or feminine in gender. The majority of these ambiguous cases occur most often with a masculine pronoun, causing them to have higher masculine than feminine probabilities in the statistical counts. Indeed, the classifier performs better by accepting masculine and rejecting feminine gender in these cases. Naturally, this is sometimes incorrect, the noun really is feminine, and such an error will result in a false negative for the feminine classifier and a false positive for the masculine one, reducing female recall and male precision, respectively.

Tests in Table 4.1 used both the corpus-based and web-mined features for classification. It is interesting to see which of these two approaches performs better individually. We train a classifier using only the parsed-corpus features and then only the web-mined features and compare performance (Table 4.2). We see that the web-mined approach can outperform the corpus-based features. Although mining the web is noisier, it has far greater coverage. Yet notice that on their own, both the web-mined at 90.4% F-Score and the corpus-based at 85.4% F-Score are eclipsed by the full classifier, at 92.2% F-Score (Table 4.1). Thus both web-mined and corpus-based features make a non-overlapping contribution; using them both results in better performance than choosing one method individually.

To put our results into perspective, we collected gender-guessing performance scores for human noun gender guessers. Three native English-speaking grad students were presented with the noun list and asked to assign the most likely gender to each token. None of the students achieved scores as high as our full gender classifier, and the average performance was only 88.8% microaveraged F-Score (Table 4.2). Thus we see that human *a priori* noun gender knowledge is on a level with our probabilistic information. To the extent this guides humans in pronoun resolution, we should expect our information to be competitive. Yet humans can achieve perfect performance, while accurate

Table 4.3: Simple Anaphora Resolution System Accuracy

| From preceding nouns, choose | Rate(%) |
|---|---|
| Most recent noun | 26.0 |
| Most recent noun without gender mismatch | 30.8 |
| Most recent noun accepted by SVM gender classifier | 59.4 |

automated anaphora resolution systems remain a definite challenge (as we verify in the following section). It is thus clear that gender information can only help up to a certain point in resolution tasks; contextual clues, entity-modelling and other, as yet unknown mechanisms must significantly compliment the background noun gender knowledge used in human pronoun resolution.

It is interesting to wonder, what is the best possible performance for a system that makes independent and consistent gender decisions? As mentioned above, a word can assume different genders in different contexts, leading to unavoidable errors for a gender classifier. We determined the highest performance possible on the classification task by assigning the correct gender to words with only one gender in the test data, and choosing the most frequently occurring gender for words which are antecedents of pronouns of differing genders. For example, the noun "agency" is neutral three times in the test set, but four times links with the plural "they" (e.g., "the agency said they will..."). Our optimum classifier assigns "agency" a plural gender and scores unavoidable false negatives on the neutral instances. Using this system, we see that the best any system can do on our test set is an F-score of about 99%.

## 4.6 Anaphora Resolution with Learned Gender Information

In the previous section we saw good performance of our probabilistic gender information in predicting the gender of a list of noun tokens. We now turn to the ultimate task: integrating this information within various full anaphora resolution strategies and measuring the improvement in performance. We describe the various systems in this section and provide the results on the test portion of our labelled data in Tables 4.3 and 4.4.

All of the systems outlined below resolve an anaphor by incrementally visiting preceding nouns and determining whether each one might be coreferent with the anaphor to be resolved. To sucessfully resolve an anaphor, the system must correctly skip over intervening noun phrases that are not antecedents and eventually accept a noun that is indeed an antecedent.

We begin with a baseline approach and then test enhancements to this framework. The baseline strategy always picks the most recent noun – the first noun it comes to when it moves backward incrementally. This unsophisticated approach correctly resolves only 26% of the pronouns in our test set (Table 4.3). Next, we incorporate gender constraints for nouns where the gender is available explicitly in the text, as explained above. This results in a modest improvement of around 5%, to 31%. We now compare this standard gender approach to an approach that uses the SVM gender

Table 4.4: Machine Learned Anaphora Resolution Accuracy

| Full-featured SVM Anaphora Resolution | Rate(%) |
|---|---|
| Without probabilistic gender information | 63.2 |
| Using probabilistic gender information | 73.3 |

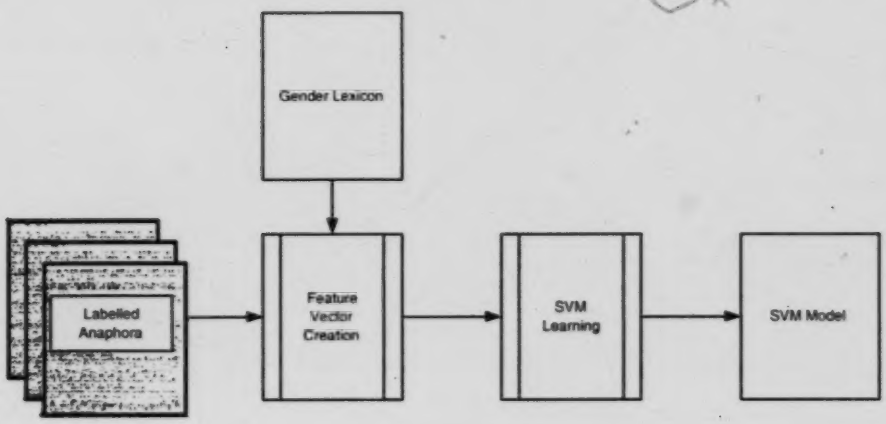



Figure 4.2: SVM anaphora resolution learning.

classifiers learned in Section 4.5. Recall, these classifiers return whether a noun is likely to be of a particular gender based on the mined gender pattern counts. Using these classifiers to reject preceding nouns that don't match the given pronoun's gender (along with rejections made by the standard gender constraints) is able to improve results substantially, up to 59%.

We next develop a more advanced anaphora resolution strategy. As discussed throughout this thesis, there are many features beyond gender information that can assist in anaphora resolution. The frequency of the noun, its distance, its grammatical position in the sentence and the parallelism between its parent and the pronoun's parent are all useful factors to consider when deciding if the two entities are coreferent. Our advanced strategy is to assign each of these features to a dimension in feature space, determine the feature signature for each potential noun-pronoun resolution decision, and use a machine learned classifier to make the coreference classification. The full feature set is provided in [7] and includes the gender features provided by our *Beta* distribution modelling.

Before determining the feature values, we first link nouns based on string match; that is, we establish coreference between different occurrences of identical nouns in the text. This allows us to determine the frequency of nouns in the text, and to send any gender information gathered from one instance of a noun to all other occurrences of that noun in the text. To create the training set, we adopt the procedure of Soon et al. [69] described in Section 2.3.1. This process provides a set with 1251 positive examples and 2909 negative examples as our training data. The SVM is trained on the training portion of our ANC data, and, like the simpler approaches described above, tested

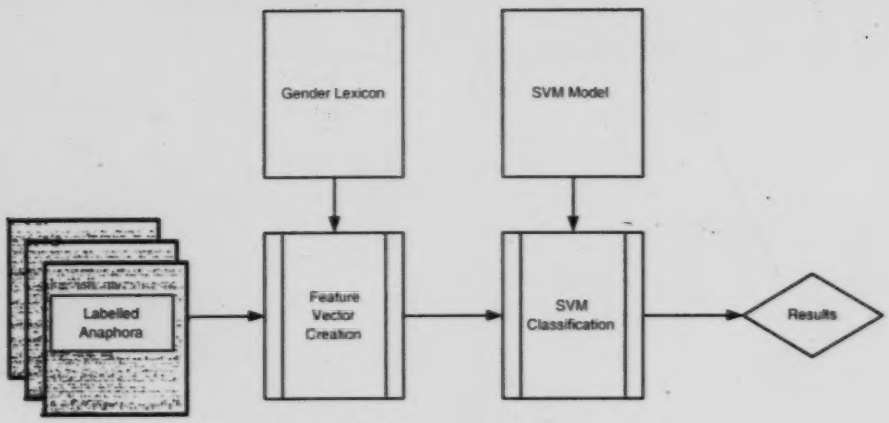



Figure 4.3: SVM anaphora resolution classification.

on the test portion by incrementally inspecting nouns preceding a pronoun until the first pronoun-antecedent match is classified as coreferent.

Figure 4.2 shows the learning process, where the gender information, along with the annotated training data, are used to build an SVM model. This model then enables classification on unseen testing data, depicted in Figure 4.3. We achieve an anaphora resolution accuracy of of 73.3% using this approach (Table 4.4). To determine the importance of the gender information to this score, we repeat the process but with the features corresponding to probabilistic gender information removed. The gender-impaired system only reaches a performance of 63.2%. Thus even for multi-featured, knowledge-rich, machine-learned anaphora resolution strategies, probabilistic gender information can provide as much as a 10% improvement in performance.

## 4.7 Conclusion

This Chapter has detailed an approach to anaphora resolution using probabilistic gender information extracted automatically from unlabelled text. In both baseline and sophisticated anaphora resolution strategies, the incorporation of probabilistic gender information enabled significant gains in performance. In addition, this information was shown to enable machine-learned classifiers to exceed human performance at noun gender-guessing tasks.

Because our probabilistic information directly depends on the quantity and quality of available text, the use of larger text repositories and continued growth of the world wide web can both improve our gender data.

There are other strongly-coreferent patterns that might be used to extract gender information from parsed corpora. For example, we show in Chapter 6 how cataphoric pronouns can be resolved with great accuracy. These resolutions could also provide instances for bootstrapping overall gender

knowledge.

However, it might be the case that bootstrapping gender knowledge from manually-defined patterns is not the most effective method to extract the information. In Chapter 5, we develop a technique for automatically determining which paths in text are highly coreferent and which ones are not coreferent. We could use the highly coreferent paths in text as our base instances, and bootstrap overall gender information from these putative resolutions.

As well, in Chapter 7, we develop a way to mine gender information from all pronoun occurrences in text. The probability of a noun's gender arises from the frequency of that noun's presence on antecedent lists of gendered pronouns. In a comparison test with the above developed system, we show that a resolution system with gender determined in this manner is competitive with the parsed-corpus/web-mined gender-driven system outlined above.

# Chapter 5

# Learning Antireflexive Coreference Relations

## 5.1 Introduction

We have seen in previous chapters how anaphora resolution systems employ a number of filters and constraints to decide whether two entities in text are coreferent. Often these filters are sufficient to block coreference, but in many difficult cases, traditional constraints do not apply, and world-knowledge is seemingly needed to decide if the two expressions can corefer. We present an algorithm that learns coreference-blocking relations from parsed text by identifying and collecting syntactic paths with non-coreferent terminal pronouns.

If a syntactic relation tends to have pronouns of the same gender/case/number in its terminal positions, it can be assumed that this relation allows coreference between its terminal entities. If this relation always has non-coreferent terminal pronouns – that is, ones of different case/gender/number, then there is a statistical bias against coreference along these paths. We call these non-coreferent paths *antireflexive relations*: an entity cannot relate to itself via this relation. Our algorithm learns all antireflexive paths in text, and is experimentally validated in a novel possessive-pronoun resolution module.

This is another example of bootstrapping to extract coreference resolution information. We use the simple but frequent cases of pronoun-to-pronoun syntactic relations to gather information for the general noun-to-pronoun path case. Information we learn in the simpler case bootstraps information that helps when resolving pronouns in unseen cases.

We motivate the problem and discuss related work in the following section. In Section 5.3 we describe the algorithm that learns antireflexive relations from parsed text. In Section 5.4 the algorithm is used to extract paths that block coreference involving possessive pronouns. In Section 5.5 we discuss the coverage of our method and present some experimental results. We show good performance, and extend coverage using an automatically-generated thesaurus. Important ideas for future work are discussed in the conclusion, Section 5.6.

## 5.2 Antireflexive Relations

Consider the following two expressions:

(1) John needs his friend.

(2) John needs his support.

In Sentence 1, "John" and "his" corefer. In Sentence 2, "his" refers to some other, perhaps previously evoked entity. Traditional coreference resolution systems are not designed to distinguish between the above cases; they lack the specific world-knowledge required in the second instance, that a person does not typically need his own support. We would say that the syntactic path between "John" and "his" in Sentence 2 forms an *antireflexive* relation. Nouns connected by these paths usually refer to distinct entities. Our goal is to learn antireflexive relations from text and utilize knowledge of them in coreference resolution.

As we have seen, coreference resolution is generally approached as a pairwise classification task, where various constraints and preferences are used to determine whether two expressions corefer. Coreference is typically only allowed between nouns matching in gender and number, and not violating any intrasentential syntactic constraints, such as Principles A, B, and C of Government and Binding Theory (See Section 2.1 and [23]). Recall that Lappin and Leass, for example, chose one antecedent for each pronoun by first applying the gender, number and syntactic filters to the set of candidate noun phrases, and then scoring the remaining candidates [38]. Machine learning approaches, meanwhile, apply constraints implicitly as decision nodes on a decision tree [2, 69].

Thus when previous systems handle cases like Sentence 1 and Sentence 2, where no disagreement or syntactic violation occurs between the expressions, coreference is determined by the weighting of features or by the learned decisions of the coreference resolution strategy. Without the knowledge that Sentence 2 represents an antireflexive relation, a resolution process would resolve the "his" pronouns in Sentence 1 and Sentence 2 the same way.

Most of the related work discussed earlier has emphasized the development of new strategies for combining well known and reliable constraints and preferences [51, 59], not the development of new factors themselves. On the other hand, we did see in Section 2.3.2 the recruitment of world-knowledge to provide new constraints in terms of interchangeability of two coreferent expressions by Dagan et al. [15]. This approach uses statistics to assess the validity of swapping the potentially coreferent entities. We use statistics to assess the validity of coreference along the path connecting the expressions. We learn these new constraints for difficult coreference instances unsupervised from text, without recourse to textual cohesion or coherence measures [25].

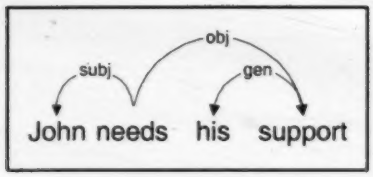



Figure 5.1: Example Dependency Tree.

## 5.3 Antireflexive Relation Extraction Algorithm

We define the *relation* between two potentially coreferent entities as the path in the parse tree between the two entities. We use the structure induced by Minipar to build the parse tree. We represent the parse tree of Sentence 2 in Figure 5.1.

A given path is said to be an antireflexive relation if it generally joins items that do not corefer. The path from "John" to "his" in Figure 5.1 is one such path. In fact, for any subject, A, joined to any genitive modifier, B, through A←subj–needs–obj→support–gen→B, A and B would generally refer to distinct entities. Our algorithm finds these relations by counting the number of times they occur with terminals that are either "likely coreferent" or "non-coreferent." In the simplest version, these special cases can be determined when the terminals are pronouns. We partition pronouns into groups of matching gender, number, and person; for example, the first person singular group would contain "I," "me," "mine," and "myself." If the two terminal pronouns are from the same group, coreference along the path is likely. If they are from different groups, like "I" and "his," then they are non-coreferent.

To formalize, let P be a path in a parse tree, and let A and B be the terminal nodes of the path. To mitigate data sparsity, P can be taken with the root form of the verbs and nouns in the path. All modifiers not on the direct path, such as adjectives, determiners and adverbs, are not considered. We extract all paths in a parsed corpus. Let $N_1$ be the number of times A and B are from the same pronoun group, and let $N_2$ be the number of times A and B are from differing groups. In subsequent application of the learned filters, a given path can be defined as antireflexive if the following conditions hold:

$$N_2 > C_1 \quad \text{(i)}$$

$$\frac{N_2}{N_1 + N_2} > C_2 \quad \text{(ii)}$$

$C_1$ is the minimum frequency of non-coreference needed to enforce the coreference blocking. $C_2$ is the minimum proportion of non-coreference. These thresholds can be learned empirically, or, alternatively, the frequency and proportion of non-coreference could be included as features in a machine learned coreference resolution system.

Our antireflexive relation extraction algorithm is similar to an algorithm used to discover inference rules, or paraphrases, by computing similarity between paths in text [44]. In that algorithm,

the fillers of the beginning and end nodes in the paths are collected, and two paths are said to be similar (and thus likely paraphrases of each other) when they have similar terminals (thus the paths occur with a similar *distribution*). Our work, on the other hand, does not need to store the fillers themselves, only whether they are from the same pronoun group or not. Different paths are not compared in any way – each path is individually deemed to be or not to be an antireflexive relation.

We illustrate our algorithm in the following section for a specific kind of path: one that joins a subject noun phrase with a possessive pronoun.

## 5.4 Antireflexive Possessive Pronoun Relations

We apply the algorithm to paths where a subject and a possessive modifier are linked through a verb or a prepositional phrase. In traditional coreference resolution, applying syntactic constraints to possessive pronoun relationships does not filter preceding noun candidates. In Binding Theory terminology, possessive pronouns are the subjects of their own governing category (governed by the noun they modify), and are thus free to be bound by preceding noun phrases [23]. Thus an approach that eliminates preceding noun candidates for possessives fills a major void in anaphora (pronoun) resolution.

For the current work, patterns are extracted by parsing about 700 million words of text, mostly news articles. The AQUAINT corpus [73], the Reuters corpus [66], and the Associated Press and Wall Street Journal sections of the TIPSTER corpus [26] are used for path extraction, while the San Jose Mercury News section of TIPSTER is used for testing (Section 5.5).

### 5.4.1 Preposition-Path Filters

We first extract antireflexive paths matching the following three patterns:

- **Preposition possessive-pronoun noun, pronoun verb...** e.g. With their help, it worked.
- **Pronoun verb preposition possessive-pronoun noun...** e.g. He went with his mother.
- **Pronoun verb object preposition possessive-pronoun noun...** e.g. She brought the toy from her room.

To mitigate data sparsity, we ignore the verb part of the path; paths with equivalent preposition and noun portions are grouped together. This sparsity reduction comes at the cost of a loss of precision when applying the filters (Section 5.5).

We can then use these cases in situations with regular nouns in the place of the subject pronoun, to either accept or rule-out coreference between the possessive pronoun and subject noun. For example, in which of the following is linking "his" to "John" acceptable?

- "With his brush, John painted."

Table 5.1: Most frequent antireflexive preposition-noun relations (greater than 75% proportion pronouns non-coreferent). Example non-coreferent entities italicized.

| Pattern | Freq: | Example |
|---|---|---|
| in-*pos*-interest: | 790 | *The acquisition* was in *its* best interest. |
| in-*pos*-opinion: | 525 | In *his* opinion, *John* is a liar. |
| in-*pos*-view: | 378 | *Mary* is a good fit in *her* view. |
| in-*pos*-face: | 289 | The *firecracker* exploded in *her* face. |
| for-*pos*-support: | 195 | *John* thanked him for *his* support. |
| out of-*pos*-hand: | 122 | The *matter* was now out of *its* hands. |
| to-*pos*-attention: | 108 | The *problem* was brought to *its* attention. |
| at-*pos*-word: | 105 | *John* took him at *his* word. |
| in-*pos*-shoe: | 103 | *Mary* wouldn't want to be in *her* shoes. |
| after-*pos*-death: | 90 | After *his* death, *John* took over. |
| with-*pos*-help: | 64 | *John* did it with *his* help. |
| to-*pos*-eye: | 59 | The *losses* brought tears to *their* eyes. |
| of-*pos*-death: | 54 | *John* heard of *his* death. |

- "With his help, John painted."

Given a subject-noun/possessive pronoun in this noun-preposition pattern, we rule out coreference (i.e. label the path antireflexive) if the proportion of non-coreferent pronouns in the collected paths is strictly greater than 75% (condition (ii)) and the number of non-coreferent occurrences is greater than 7 (condition (i)). Similarly, we say a pattern is *not* antireflexive if it occurs more than 7 times with greater than 25% of terminals of the same group (this working definition is used in our experiments in Section 5.5). If it's below our frequency threshold, we allow coreference by default. For example:

- with-*pos*-brush: 2 occurrences, both of same group, allow coreference by default.
- with-*pos*-help: Of 67 occurrences observed, 64 non-coreferent; proportion is 96%: do not allow coreference.

As expected, the algorithm extracts paths with terminal pronouns of the same group much more frequently than paths with non-coreferent terminals. For example, of the 1103 occurrences of the pattern by-*pos*-wife ("John stood by his wife"), 99.3% are of the same group. For patterns with such a high proportion of matching pronouns, a coreference resolution system would do well to resolve the possessive pronoun to the subject noun by default.

The most frequent preposition-noun combinations with non-coreferent pronouns capture instances we intuitively agree are antireflexive (Table 5.1). All the most frequent examples seem legitimate. They primarily reflect common expressions with non-coreferent *pronouns* ("it was in her best interest to..."), not necessarily the most non-coreferent paths with *nominal* subjects. Nevertheless, this information does indeed prevent us from making silly coreference decisions. How can

Table 5.2: Most frequent antireflexive verb-noun relations (greater than 75% proportion pronouns non-coreferent). Example non-coreferent entities italicized.

| Pattern | Freq: | Example |
|---|---|---|
| break-*pos*-heart | 248 | *Jamie* broke *his* heart. |
| change-*pos*-life | 189 | The *experience* changed *his* life. |
| need-*pos*-help | 182 | *John* needs *his* help. |
| cross-*pos*-mind | 158 | His *son* crossed *his* mind. |
| save-*pos*-life | 140 | *She* saved *her* life. |
| know-*pos*-name | 123 | *Mary* knows *her* name. |
| recall-*pos*-life | 122 | The *novel* recalls *his* early life. |
| feel-*pos*-pain | 107 | The *president* feels *her* pain. |
| see-*pos*-face | 100 | *John* saw *his* face in the crowd. |
| have-*pos*-support | 93 | *Mary* has *her* full support. |
| shake-*pos*-hand | 91 | *He* shook *his* hand vigorously. |
| hear-*pos*-voice | 87 | *John* heard *his* voice on the phone. |
| reach-*pos*-desk | 76 | The *files* reached *their* desk. |

someone take over when they're dead? How could they hear of their own death? Substituting in different nouns or prepositions would make the coreference acceptable:

- Mary is a good fit in her *job*.
- *Before* his death, John took over.

Thus our approach gathers the specific word combinations that are antireflexive, without filtering similar yet valid sentences. Naturally, exceptions to the antireflexivity of some of these patterns can be found (especially when less-frequent senses of the path nouns are used); our patterns represent general trends only.

### 5.4.2 Verb-Path Filters

We next collect instances of the following pattern:

- **Pronoun verb possessive-pronoun noun.** e.g. He visited his wife.

As with the preposition-noun combinations, for each verb-noun pair, we count the number of paths with pronouns in the same group (same gender, number, and person) and in different groups (different gender, number, or person).

Here the entire path is included, so there is less variability in the semantic content of a given pattern. We can confidently make antireflexivity decisions with fewer extracted examples of the paths than we could in the prepositional case above. We thus relax the frequency threshold, preventing coreference if the number of non-coreferents observed is greater than four (condition (i)), but keeping the proportion of non-coreference cutoff at 75% (condition (ii)). For example:

- "John welcomed his opponent." welcome-*pos*-opponent: 0 in different group, 1 in same group: allow coreference by default
- "John welcomed his suggestion." welcome-*pos*-suggestion: 7 in different groups, 0 in same: do not allow coreference

The most frequent cases with a high proportion of pronouns in the same group once more represent situations where coreference is almost automatic between the two nouns. The top three most frequent are change-*pos*-mind, do-*pos*-job, make-*pos*-way, having pronoun-match frequencies of 1241, 1069 and 783, respectively.

Table 5.2 again reflects our intuition about antireflexive paths. Our patterns eliminate coreferences that would be nonsensical, infrequent, or obvious. You don't break your own heart, rarely shake your own hand, and it's usually not necessary to state that someone knows their own name. On the other hand, we do find the pattern recall-*pos*-life a little dubious. It owes its frequency to the repeated use of "it recalls his ..." in a small section of our corpus.

As with the prepositional paths, different nouns or verbs would make the links acceptable:

- Jamie broke his *watch*.
- Mary *gave* her name.

We next show these learned patterns can indeed improve the performance of general coreference resolution applications.

## 5.5 Experimental Results

We perform our experiments on the SJM corpus, which is disjoint from the data on which we applied our extraction algorithm (Section 5.4). Parameters $C_1$ and $C_2$ are loosely set based on performance on a separate development portion. Although we focus on specific prepositional and verbal possessive pronoun patterns, these patterns nevertheless cover a significant portion of all pronouns. Of the first 149819 possessive pronouns extracted from the SJM corpus, 49022 match our patterns, or roughly one third.

A key question is: in what proportion of our patterns, in general, does the possessive pronoun resolve to the subject of the pattern? We examined the first 500 occurrences of our pattern in the corpus, and manually decided whether the possessive and subject corefer. Of the 500, 393 corefer while 107 do not. Thus, a system could achieve 80% performance on these cases by simply resolving the possessive to the subject, irrespective of gender, number, frequency or other considerations. In fact, one can do significantly better than 80% by applying gender and number constraints, but as mentioned above, there will always remain the difficult cases where the usual filters do not apply. These are the cases we would like our system to handle.

Table 5.3: Results for Antireflexive Paths.

| Type | Correct | Incorrect | Rate |
|---|---|---|---|
| Prep | 129 | 54 | 70.5% |
| Verb | 384 | 45 | 89.5% |
| Total | 513 | 99 | 83.8% |

Table 5.4: Results for Similar Antireflexive Paths.

| Type | Correct | Incorrect | Rate |
|---|---|---|---|
| Prep-similars | 49 | 30 | 62.0% |
| Verb-similars | 197 | 46 | 81.1% |
| New Total | 759 | 175 | 81.3% |

We applied our antireflexivity criteria to the 49022 matches, and found 612 instances to be antireflexive. Thus, 1.2% of coreference resolutions are blocked by our approach, supposedly representing a difficult portion of the 20% of cases where coreference does not occur. We then manually went through each of the 612 blocks and decided whether the blocked entities were indeed non-coreferent (correct), or whether the blocked entities do corefer (incorrect) (Table 5.3). For the most part, coreference is correctly blocked. The verb patterns perform significantly better than the prepositional ones, likely because the verb patterns represent a complete path, while, as mentioned above, we ignore the verb portion of the prepositional paths. Precision is indeed impaired by this abstraction.

Overall, these results are encouraging. As mentioned above, 80% of possessives resolve to the subject, yet our verb-path antireflexivity detection, for example, identifies a subset where only 10% of subjects and possessives corefer. Coreference resolution systems would benefit from this information.

Next, we investigate whether we can increase the coverage of our algorithm by considering words similar to the nodes in our paths. For example, need-*pos*-support is identified, but need-*pos*-backing and need-*pos*-assistance are below the frequency threshold, and thus not considered antireflexive. Could these infrequent paths also be blocked by virtue of their association with an antireflexive path?

We use an automatically constructed thesaurus developed by Dekang Lin, based on the distributional pattern of words [41]. This thesaurus captures the word collections we desire: the most similar nouns to "support" are "backing," "assistance," and "help." Similarity does not imply synonymy; "father" and "mother," for example, are very similar words in this thesaurus, but are not synonymous.

For each word in the noun node of low-frequency paths, we look up the nine most similar nouns from the thesaurus, and thereby create nine new metapaths for consideration. We make decisions based on these paths conservatively: if one or more of the paths meet our antireflexive criteria, and

none of them are found to be not antireflexive, then we enforce coreference blocking for the original pattern.

Use of the similar paths increases the coverage of our approach to about 2% of matching preposition and verb patterns. The performance on the similar-path blocked patterns runs at 81% for verb paths and 62% for prepositions, reducing overall precision to 81.3% (Table 5.4). We thus observe an expected trade-off between coverage and precision.

The coverage versus precision issue can also be addressed by lowering the thresholds in (i) and (ii) and considering more paths as non-coreferent. When we lower the non-coreferent threshold (ii) from 75% to 50%, including similar paths, our coverage increases to 3.9%, but overall precision decreases to approximately 68% (80% for verb-noun patterns, 59% for prep-nouns). Whether the increased coverage compensates for the lower precision, or where the line between coverage and precision should indeed be drawn, are questions that can only be addressed when antireflexive path information is evaluated within a full coreference resolution system. That is the platform in which the optimal configuration of parameters must be obtained.

## 5.6 Conclusion

In this chapter, we have proposed a new source of constraints for coreference resolution: antireflexive relations. We demonstrated our algorithm using possessive pronoun patterns, and showed that it extracts intuitive and important coreference-blocking relationships. We validated the scope and performance of these filters experimentally, and showed better performance with a complete-path verb filter than with the partial-path prepositional pattern.

We are currently developing tools to extract all antireflexive paths, rather than selecting common important paths like the preposition-possessive and verb-possessive experiments done above. There are many exciting coreference-blocking relations remaining to be discovered. Furthermore, we will increase the coverage in four ways: by using more data, expanding the definition of similar paths, grouping together equivalent but syntactically-different paths, and using nominal expressions (with known gender and number) as terminals in the extraction algorithm. We might also investigate using the web to collect the non-coreferent proportion information. For a given path, we can decide antireflexivity by counting the number of pages returned by a search engine when the path terminals are from the same or different pronoun groups.

Our antireflexive relation extraction algorithm provides a powerful new tool for discourse interpretation researchers. It is broadly-applicable and easily incorporated into coreference resolution systems, with little cost and much room for resolution performance improvement.

# Chapter 6

# Non-Anaphoric Pronouns

Not every pronoun in text refers anaphorically to a preceding noun phrase. There are a frequent number of difficult cases that require special attention, including pronouns that are:

(1) Pleonastic: pronouns that have a grammatical function but do not reference an entity. E.g. "It is important to observe it is raining."
(2) Cataphora: pronouns that reference a future noun phrase. E.g. "In his speech, the president praised the workers."
(3) Non-noun referential: pronouns that refer to a verb phrase, sentence, or implicit concept. E.g. "John told Mary they should buy a car."

Because our bootstrapped approaches to anaphora resolution require extracting information from unlabelled pronouns in text, the presence of non-anaphoric pronouns might pollute our results with meaningless or potentially erroneous coreference connections. Therefore, we develop ways to automatically identify and handle such cases.

Specifically, the modules described in this chapter are used prior to learning in our Expectation Maximization approach to pronoun resolution, described in Chapter 7. We filter pleonastics from the candidate lists and from being instances for the learner, and add future nouns as candidates in the case of cataphora. This ensures more robust data is included in the iterative determination of our probability distributions.

In this chapter, we show how pleonastics can be identified syntactically using an extension of the detector developed by Lappin and Leass [38]. Roughly 7% of all pronouns in our test data are pleonastic.

For cataphora, we adopt a different approach than Lappin and Leass, who always include all future nouns from the current sentence as candidates, with a constant penalty added to possible cataphoric resolutions [38]. We are able to identify cataphora constructions automatically using the parser output, and include future nouns in the candidate list in these cases. We will describe this process in detail. Our cataphora module identifies 1.4% of pronouns in the AQUAINT test data (the

test set for the work in Chapter 7) to be cataphoric, and in each instance this identification is correct (100% precision). Further experimental results are presented in Section 6.2.6.

In this chapter, we first explain our work in handling pleonastics, while covering some related approaches, and then we discuss our novel cataphoric pronoun handler.

We are not aware of any ways to avoid the errors caused by non-noun referential pronouns; this will be the subject of future research. The unavoidable errors for these implicit pronouns, occurring roughly 4% of the time in the AQUAINT test data, are included in the final results of the Expectation Maximization approach to pronoun resolution described in Chapter 7. Difficulties with such cases are not unique to computational approaches; Chomsky discusses Binding Theory issues for sentences like Sentence 3 in Chapter 5 of [13].

## 6.1 Pleonastic Pronouns

Pleonastic pronouns are always represented in text by the two-letter pronoun, "it." Pleonastics often appear as the first element in the sentence, but can also show up as a dummy object to a verb:

(4) It appears John is late.

(5) It turned out the money was counterfeit.

(6) They've really blown it – big time.

(7) Aw, darn it!

We describe previous approaches to identify pleonastics, and describe the method we implemented to filter pleonastics in our Expectation Maximization approach to pronoun resolution.

### 6.1.1 Previous Approaches

There are a variety of previous approaches to identifying pleonastic pronouns in text. These include both rule-based pattern-matching approaches and machine-learned approaches.

#### Lappin & Leass

Lappin & Leass report that their work is the first computational treatment of pronominal anaphora resolution to address the problem of pleonastic pronouns [38]. A combination of lexical and syntactic filters are used. First of all, a class of so-called "modal adjectives" are specified, including the following:

```
necessary
good
economical
possible
useful
easy
certain
advisable
desirable
```

etc. These are used to identify pleonastic pronouns in the following constructions:

- It is **Modaladj** that **Sentence**
- It is **Modaladj** (for **Noun Phrase**) to **Verb Phrase**
- **Noun Phrase** makes/finds it **Modaladj** (for **Noun Phrase**) to **Verb Phrase**

Secondly, a group of so-called "cognitive verbs" is specified:

```
recommend
think
believe
know
anticipate
assume
expect
```

These are used to identify pleonastic constructions of the form:

- It is **Cogv-ed** that **Sentence**

Other pleonastic uses of "it" are identified if they match:

- It seems / appears / means / follows that **Sentence**
- It is time to **Verb Phrase**
- It is thanks to **Noun Phrase** that **Sentence**

Kennedy and Boguraev present a modified and extended version of Lappin & Leass's work that does not require full syntactic parsing. They report, however, that the 306 anaphoric pronouns on which they state results do not include 30 instances where the pleonastic detector (referred to in their paper as the "expletive" detector) failed to identify the pleonastic pronoun. It is not clear whether this large number of improperly handled pronouns (which, if included in the results, would have reduced their accuracy from 75% to about 68%), is primarily because of only shallow parsing of the input text. If that were the case, then this would present a major drawback of attacking pronoun resolution without a parser.

**Evans**

Evans identifies pleonastics with a machine learning approach that classifies all instances of the word "it" [17]. Feature vectors are built for all occurrences of "it" in text, and supervised machine learning is used to learn a classifier. Evans sought to consider all cases of non-anaphoricity simultaneously, with pleonasticity as well as cataphoric, or non-noun referential reference also being considered. Thirty-five features are used in classification – including positional information, parts of speech of surrounding elements, proximity of complementizers, etc. Evans reports results on the subtask of

Table 6.1: Taxonomy of definite noun phrases

| | | | |
|---|---|---|---|
| Definite Noun Phrases: | 1) | Referential | ("the president") |
| | 2) | Existential | |
| | | 2a) Associative | ("the score") |
| | | 2b) Independent | |
| | | 2bi) Semantic | ("the FBI") |
| | | 2bii) Syntactic | ("the organization which...") |

only deciding whether the pronoun "it" is pleonastic or not, and achieves a precision of 73.38% and a recall of 69.25%.

This system is used as a preprocessing step in Mitkov et al.'s fully-automatic system [53]. In the evaluation section of [53], the classification accuracy for *it* is stated as 85.54%. Classification accuracy means the total number of "it"-pronouns correctly classified as either anaphoric or pleonastic, as a percentage of the total number of *it* occurrences. In our opinion, accuracy is not a useful measure for this task. Since total counts of anaphoric and pleonastic pronouns are given in one of their tables, we can see that had the classifier simply assigned the "anaphoric" category to each occurrence, the classification accuracy would have been 82.0%. In general, precision and recall measures are more informative for sparse classification.

**Non-anaphoricity in coreference resolution**

Non-anaphoricity is an important consideration for all noun phrases in general coreference resolution, beyond only identifying pleonastic pronouns. This is because most nouns in text are not involved in coreference relationships, and knowing exactly which ones do not refer to previous entities would enable a major boost in coreference resolution performance [24].

For example, if we see "the president" in text, we need it resolved to find out exactly which president we are referring to. But if we see "the White House," then we are already expected to have a strong notion of what this phrase embodies. How can we determine which phrases need to be resolved and which ones do not?

Bean and Riloff [4] look at finding definite noun phrases (ones modified by a definite article) that are non-anaphoric by virtue of them being part of the general world-knowledge of the reader.

They devise a general taxonomy of noun phrases, the specification of which would allow non-anaphoric noun phrases to be immediately identified in a coreference resolution system (Table 6.1). Referential phrases are anaphoric in the document. Associate phrases are known to the reader, but what exactly they refer to depends on the context of the article. The paper suggests perhaps 10% of existential NPs are associatively existential. Semantic independent phrases rely on the world-knowledge of the reader to decide on the referent, while syntactic phrases are defined in the remainder of the sentence in question.

How do they decide which nouns fall into which categories? Bean and Riloff [4] have vari-

ous syntactic heuristics for identifying some of them, but they also use a unique method looking at discourse-initial noun phrases in text (a method that we later independently proposed to mine pleonastics and cataphora). If a noun occurs in the first sentence of a document, and it's not syntactically independent, then it's semantically independent.

They also perform a number of operations to extract domain-specific knowledge, such as making inferences from nouns that occur frequently in text as definite but rarely as indefinite. For example, we see the phrases "a FBI" and "a White House" much less often than their definite counterparts, implying these phrases are semantically independent.

Although their discourse-initial noun extraction approach applies to general noun phrases, it is clear that it may also be applicable to pleonastic pronouns. Of course, discourse-initial pronouns might also be cataphoric rather than exclusively pleonastic, precluding pure application of the extraction approach above. But fortunately, in the following section we develop a cataphora detection module. We could use this module by extracting all discourse-initial-noun occurrences of "it," and filter those that are cataphoric. The remaining cases would be largely pleonastic. Thus a variety of pleonastic constructions could be obtained automatically, and used elsewhere in text for pleonastic detection.

It is also worth mentioning that developing pleonastic pronoun detectors separately from non-anaphoric noun phrase detection in coreference resolution would be beneficial. Ng and Cardie [58] choose to handle pleonastics as a subset of general non-anaphoric noun phrase identification. Like Evans, they provide a number of features to a classifier and learn decision trees for anaphoricity (as well as rule-induction classification with RIPPER [14]). Using the same set of features for common nouns and pronouns usually learns decision trees that do not favour pronouns. In the top portion of the decision tree provided as a figure in their paper, we see that in all the depicted branches of the tree involving pronouns, anaphoricity determination for pronouns basically amounts to always classifying each pronoun as anaphoric. In all tabular results, recall performance on pronoun anaphoricity is between 10 and 20%, with corresponding low F-measure scores. Thus it appears that a custom non-anaphoric noun phrase detector would have been beneficial, whose output itself might have been included as a feature for the general noun phrase anaphoricity determination.

### 6.1.2 Our Approach

Lappin & Leass [38] suggest that it could be argued that the identification of pleonastics is really a job for the syntactic and semantic analysis (parsing, named-entity-recognition) that precede anaphora resolution. This is somewhat true of the parser that we use, Minipar. We are able to identify certain pleonastic constructions automatically because Minipar identifies these cases with a "Subj" category label (e.g. "It (Subj) appears that...").

For ones not picked up by the parser, we use the lexico-syntactic pattern-matching procedure of Lappin & Leass [38] to identify the remainder. We build our own lists of modal adjectives and

Table 6.2: Pleonastic detection confusion matrix

| | | Classification | | |
|---|---|---|---|---|
| | | Pleonastic | Anaphoric | Total |
| Truth | Pleonastic | 51 | 31 | 82 |
| | Anaphoric | 30 | 1097 | 1127 |
| | Total | 81 | 1128 | 1209 |

cognitive verbs and we implement the patterns outlined in the Lappin & Leass description above, but for a dependency-based syntax.

The named entity recognition components of Minipar are useful in identifying time-expressions – for example, to identify the time expression and hence the pleonastic in "it was midnight." Frequent special cases observed in our development set are encoded as well, especially when the pleonastic is in the object position (e.g., "darn it, blown it, overdo it, endure it," etc.)

These expressions are not always pleonastic – sometimes the pronoun does in fact refer to a previous entity. However, when we developed our pleonastic-identification system, it was for use in our Expectation Maximization approach to pronoun resolution, and therefore our primary objective was to prevent pleonastic pronouns from reaching our unsupervised learner. Even if the module was too aggressive, we could afford to miss a few truly anaphoric cases and simply incorporate more data to compensate. That is, in terms of pleonastic detection, we sought a system with high recall (few false negatives) at the expense of precision (more false positives). In the end, however, our system had a fairly equal precision-recall trade-off. We show the confusion matrix in Table 6.2 for our AQUAINT test data. Overall precision is 63% and recall is 62%. Since most pronouns are anaphoric, overall accuracy is 95%. Again, note that simply labelling each pronoun as anaphoric would have a similarly high accuracy of 93%.

## 6.2 Cataphora

### 6.2.1 Introduction

Cataphora, also sometimes called backward pronominalization [27], occurs when a pronoun is mentioned before its antecedent. Such situations have previously presented difficulties to anaphora resolution researchers, and are often either manually identified and skipped (as we do with our system in Chapter 4, above), or ignored with a subsequent loss in performance. Systems that learn coreference resolutions from unlabelled text or apply coreference decisions to previously unseen data (like our system in Chapter 7), however, need tools to deal with these cases. In this thesis we show that the majority of cataphora occur within the same sentence as their antecedent, and can, in fact, be consistently identified and resolved with great accuracy, on par with the supposedly simpler and more widely studied case of anaphora resolution. A novel component of our fully-automated system is the antireflexive possessive-pronoun relation filter developed in Chapter 5, which, although based on

simple statistics, provides meaningful recommendations for anaphoric or cataphoric coreferences.

We provide further examples of cataphora in Section 6.2.2, and discuss the scope of our work, and that of other approaches, in Section 6.2.3. In Section 6.2.4, we describe our intrasentential cataphora identification and resolution strategy, and discuss the coverage of our approach. Section 6.2.5 describes the determination of antireflexive cataphoric relations. In Section 6.2.6 we provide our experimental results, showing that our system performs at a level comparable to anaphora resolution approaches. Finally, in our conclusion we discuss avenues of future research.

### 6.2.2 Cataphora Resolution

Anaphora resolution determines which previous entity (the antecedent) a given noun phrase (the anaphor) refers to. In cataphora resolution, on the other hand, the antecedent (for lack of a better word), occurs *after* the referring entity. Consider the following excerpt from the New York Times section of the popular Question Answering corpus, AQUAINT [73]:

(8) In his eight years as chief executive of Washington Mutual, Kerry Killinger has transformed the company from a small Seattle-area savings bank and home lender to a multi-billion, coast-to-coast financial giant.

The pronoun "his" in this sentence is actually coreferent with the future entity "Kerry Killinger." Nowhere previously in the text does "Kerry Killinger" appear; in fact, this is the first sentence in the news article. Humans resolve "his" to "Kerry Killinger" easily, and process the information that "Kerry Killinger" has served for eight years as chief executive of Washington Mutual. We would like information retrieval applications to also have access to such facts.

As mentioned earlier, the phenomenon of cataphora is largely ignored by coreference resolution researchers. Summaries of anaphora resolution research simply do not address cataphora resolution [52]. Yet if coreference or anaphora resolution systems are to be applied to real-world data, no manual identification will be possible. Also, as mentioned above, in our Expectation Maximization approach to pronoun resolution we need to identify these pronouns automatically to prevent them from polluting our data (pollution would occur when we assume the antecedent occurs in the list of previous nouns, when it actually occurs later in the sentence). We want to ensure that at least one antecedent is in the candidate lists, even if we have to look forward to find it. Why suffer the performance loss caused by simply "writing-off" cataphora?

Our novel procedure for resolving within-sentence cataphora (*intrasentential* cataphora) is outlined below. When the cataphor's antecedent occurs in a future *sentence*, as in Sentence 9, from the Xinhua section of AQUAINT, it is called *intersentential* cataphora:

(9) "He conquered the 8,848-meter-high Mount Qomolangma (Mt. Everest), the world's highest peak, in 1987. He reached the North Pole in 1991. Now he has become the first person in the world to take a west-to-east route to traverse the Sahara Desert on foot.

Choi Jong-Ryul, a 38-year-old adventurer from the Republic of Korea, overcame a series of hellish challenges..."

Our system finds the future antecedent of intrasentential pronominal cataphora only.

### 6.2.3 Scope and Related Work

Our approach is to create a definition of cataphora that facilitates coreference resolution. We wish to cover a variety of related sentences of the form:

(10) "In his speech, John described his research."

(11) "Humbly accepting her award, Mary wept."

(12) "If he's not careful, Fred might lose her."

(13) "After her climb, Sarah's feet were tired."

We want to process these sentences, apart from context, and immediately determine his=John, her=Mary, he=Fred, and her=Sarah, just as any human reading these sentences is apt to do. Some researchers distinguish between "genuine cataphora" and general intrasentential cataphora: genuine cataphora meaning the referent has not yet been evoked in the discourse, i.e. the antecedent occurs nowhere previously [36, 18]. This distinction is fairly artificial, as for practical purposes, whether previous antecedents exist or not, in each of these sentences we seem to have all the information we need to determine coreference within the sentence itself. The cataphoric part of the sentence, be it the prepositional phrases in Sentences 10 and 13, the verb phrase in Sentence 11, or the subordinate clause in Sentence 12, each cannot exist on its own. As we read, we delay resolving the pronoun until we process the main clause, and make the resolution there if a plausible antecedent exists.

Researchers who manually label and ignore *genuine* cataphora in their resolution systems, but handle cases where the antecedent does exist somewhere previously, would have to look back to preceding antecedents for coreference for the potentially non-genuine intrasentential cases above. Looking backward, rather than forward in the sentence, would needlessly make the above constructions more challenging than they need to be. Why look back five or six sentences for the antecedent when it occurs in the main clause of the same sentence?

It is worth noting that most forms of cataphora can be excluded using Binding Theory [23].

(14) *She$_i$ knows that Mary$_i$ is admired.

She cannot command Mary and refer to the same referent – by Principle C, Mary must be free everywhere. On the other hand, consider the following headline from a recent issue of the *Economist*.

(15) His foreign critics need to notice that George Bush has now done what they want.

"His" does not command George Bush so coreference is possible and the sentence is grammatical. Yet is this cataphora? The spirit of a phrase like this is editorial. Although it occurs as a headline, we perceive it more as something someone would say in the middle of a discourse or conversation, when the entity of George Bush has already been asserted, and thus does not represent true cataphora.

Langacker [37] encompasses cataphora in his proposal that the antecedent of a pronoun must precede or command the pronoun (referenced in [27]). The idea is motivated by the fact that the cataphoric antecedent in Sentences 10, 11, and 12 above would all command the pronoun in various parse tree representations. Notice, however, this would not handle Sentence 13. Our algorithm does handle such cases.

Lappin & Leass handle cataphora in their anaphora resolution system [38]. They always include all the nouns from the current sentence which pass their syntactic filters (which essentially encode Binding Theory constraints) as potential antecedents, including future nouns. To encode the fact anaphora are more frequent than cataphora, they just give large penalties to antecedent scores that involve cataphora. This process, then, makes no use of the general patterns we sketch in Sentences 10, 11, 12 and 13, which show intrasentential cataphora generally only occur when there is a dependence of the cataphoric expression on a later main clause. For unsupervised learners in particular, including all the future nouns from the current sentence every time an anaphora is encountered would needlessly multiply the candidates and impair the performance.

Abraços and Lopes handle certain instances of cataphora automatically with their parser – the syntactic tree always shows the internal arguments of a verb on its right, converting cases like Sentence 10 to "John described his research in his speech," which can be resolved like regular anaphora [1]. One might be tempted to think all cataphora can be handled in this way, and the verb phrase in Sentence 11 and the subordinate clause in Sentence 12 might also be moved to the right hand side. Ultimately, though, deciding where to attach the previous fragment presents difficulties in itself, and, in fact, sometimes leads to ungrammatical constructions. Consider the following examples adapted from Mann and McPherson [46]:

(16) If she$_i$ likes Fred$_j$, Mary$_i$ will give him$_j$ a kiss.

Moving the subordinate clause to the right:

(17) Mary$_i$ will give him$_j$ a kiss if she$_i$ likes Fred$_j$.

Moving the clause now prevents the previously resolved him-Fred coreference, and is thus invalid. It will be clear that moving clauses is unnecessary as we now explain the syntax-driven approach we have developed to identify cataphora in parsed text.

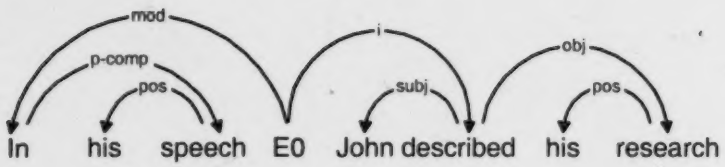



Figure 6.1: Example Cataphora Parse.

### 6.2.4 A Cataphora Resolution Strategy

**The Pattern**

As our above examples indicate, intrasentential cataphoric pronouns occur when an earlier, subordinate sentence component modifies a later sentence main clause. We used the structure induced by Minipar to identify this situation.

Minipar creates trees that generally have a top-level clausal node labelled as E0. E0 is an implicit node, that is, unlike other nodes in the parse, it has no explicit words in the sentence corresponding to it. Rather, it is a required structural feature of a Minimalist-grammar parse tree. E0 is usually the parent of the verb in the main clause, and for cataphoric constructions, it is also the parent of earlier subordinate phrases.

For example, the parse from Minipar for Sentence 10 gives the structure in Figure 6.1. Subordinating conjunctions, verbs without noun modifiers, prepositions, etc., all receive the *mod* relationship from the parser, with their parent being an E0 clause. The cataphoric pronouns are always descendants of these *mod* nodes in the parse tree, and, as mentioned the main clause's verb will also have E0 as its parent. This gives us all the information we need to identify cataphora: when a pronoun has an ancestor node (be it a preposition, verb, or subordinating conjunction) that modifies an E0 clause, and a verb in the main clause also modifies E0, then we potentially have cataphora. This is almost a complete specification, although it includes a number of cases that are clearly not cataphora. Consider:

(18) When John took his son to the hospital, Bob was concerned.

(19) According to him, Bob was concerned.

Each of these cases meets the criteria of the template outlined above. But in Sentence 18, linking "his" to the future antecedent "Bob" is incorrect – "his" already has a suitable antecedent in the cataphoric clause itself – "John." To prevent these resolutions, when an earlier, within-sentence noun is a possible candidate for a pronoun, we do not classify such pronouns as cataphora. Candidates are

nouns that pass gender, number, Binding Theory constraints, etc. This restriction will exclude some sentences that are actually cataphora, requiring these pronouns to be resolved by other means.

The problem with Sentence 19, on the other hand, is that this actually violates principle B of Binding Theory: "Bob" is in "him"'s governing category, where a pronoun must be free (not bound) [23]. To filter these situations, we do not handle pronouns without a complete governing category (i.e. ones that do not include an accessible subject) as cataphora.

**The Strategy**

Now, given a set of sentences meeting our pattern, it is noteworthy that the antecedent of the cataphoric pronoun is almost always the subject modifier of the main clause verb. As a first rule for cataphora resolution, we say the cataphoric pronoun is coreferent with the subject noun when this noun is a permissible candidate (again, permissible candidates must pass gender and number constraints with the pronouns). Two easily identifiable exceptions are given:

(20) After his death, John's achievements were celebrated.

(21) After his death, there was great mourning at his passing.

In Sentence 20, we would set "achievements" as the antecedent of "his" by default, but because "achievements" does not meet our number-agreement constraint, it is rejected. In the next example, Sentence 21, "there" is parsed as the existential subject, and is thus excluded as the antecedent. With these cases in mind, we developed the following rule for pronoun resolution:

**Cataphora Resolution Procedure**

A If the cataphor and main clause subject pass constraints, set coreference.

B Else if the cataphor and a modifier of the main subject pass constraints, set coreference

C Else if a pronoun of matching gender occurs in the main clause, set coreference

D Else, no cataphora

The algorithm in this form is adopted in the experiments below. In Chapter 7, however, we include all constraint-passing entities in the main clause on the candidate list: main subject, modifier of main subject, and matching pronouns in the main clause. In that work, a single antecedent from this list is not chosen until after several iterations of Expectation Maximization builds probabilities for all the links.

**Coverage**

One might justifiably question the coverage of a cataphora resolution system with such restrictive filters. But consider that in practice, rejected cases are not lost, but handled by an anaphora resolution

strategy, as all cataphora would have been handled without our approach. Indeed, many more cases are treated as cataphora in our approach than in past systems.

Recall that of the 2779 pronouns labelled in the *slate* section of the American National Corpus, only 16 were labelled as cataphora (Chapter 3). That is, we deemed it impossible to resolve 16 pronouns by virtue of them being cataphora. Some of these cases actually had a preceding antecedent, in addition to a future cataphoric one, it was just too far back to make resolution possible. Looking at all the labelled cataphora, we see that only 7 should be considered "genuine cataphora" – i.e. absolutely no antecedents occur in the preceding parts of the text. However, when we search for our new pattern in the same data, 39 sentences fit the pattern.

In another experiment, we gathered a rough estimate of the frequency of the more challenging *inter*sentential cataphora not handled by our system. We used the fact that generally these cataphora occur as introductory stylistic devices in the first lines of an article. For one month of New York Times news articles from AQUAINT, we output the first line of every article where the first noun is a pronoun. We then manually removed pleonastic pronouns (i.e. non-anaphoric "it" like in "it is raining"), intrasentential cataphora, and counted the number of true intersentential cases. In total, 224 intersentential cataphora were extracted. By comparison, over the same month of articles, 3301 sentences matched our cataphora pattern. There were 198,028 total pronouns extracted in the same block of text, confirming earlier statistics that show roughly 1.5% of pronouns in news text are matched by our cataphora identification pattern.

It should thus be clear that our system addresses more than what researchers previously might have considered as cataphora, and that intrasentential cataphora are the most frequent and important case requiring resolution.

### 6.2.5 Pronoun Compatibility Filters

In our first experiments using the above Cataphora Resolution Procedure, we noticed some consistent and seemingly related patterns erroneously identified as cataphora. Many phrase types like the following occurred:

(22) Under *its* leadership / stewardship / guidance / direction etc., *the company*...

When we read this phrase, we know that "its" and "company" do not corefer. In fact, it seems like leadership implicitly takes company as an argument ("under its leadership of the company, the company"), preventing coreference by Binding Theory considerations. With different noun possibilities, however, cataphora is perfectly valid ("Under its *expenses*, the company lists..."). There seems to be certain preposition-noun or verb-noun phrase combinations that generally do not link coreferent entities. This is what led us to devise the empirical method to learn the two kinds of these possessive pronoun coreference-blocking combinations, which we presented in the previous chapter (Chapter 5).

Table 6.3: Results of Cataphora Resolution Test.

| | Correct | Incorrect | Rate |
|---|---|---|---|
| XIE | 631 | 69 | 90% |
| SJM | 568 | 132 | 81% |

We test the applicability of the antireflexive filter on the cataphora task by including it in our overall Cataphora Resolution Procedure. Before looking forward for antecedents, we immediately block cataphora resolutions that are deemed antireflexive.

To configure the antireflexive algorithm for this cataphoric use, we applied the antireflexive path extraction procedure on the AQUAINT data set, and set the parameters $C_1$ and $C_2$ based on cataphora resolution performance on the Xinhua portion of AQUAINT. Results are given below.

### 6.2.6 Experimental Results

**Cataphora Resolution Performance**

We tested our cataphora resolution strategy on two corpora. Our system was originally designed to help identify cataphora in AQUAINT, and, as mentioned earlier, our antireflexive filter-patterns were collected from mismatching pronouns in that corpus. Thus, we wished to test the performance of our cataphora strategy on this set, which might be thereby considered the development set for our system. We extracted 700 cataphora from the Xinhua news section of AQUAINT (XIE), and in each instance manually evaluated whether we correctly picked (or blocked) the correct antecedent (Table 6.3). Ninety percent accuracy was achieved. This text is somewhat regular, with many of the same cataphoric expressions used again and again in different contexts ("in his speech, the president of *X* said..."). This no doubt contributed to our solid performance.

A more meaningful test was conducted by applying our automated cataphora resolution strategy to unseen text, also news articles, this time from the San Jose Mercury News corpus (SJM). Here we achieved the more realistic performance of 81%, again over 700 manually evaluated example trials (Table 6.3).

Of the 132 errors observed in this evaluation, many were the result of unidentified pleonastic pronouns, bad parses, and non-cataphoric expressions that might have been blocked by better gender and number information or a more robust antireflexive possessive-pronoun resolution filter. Continuing research will determine an upper-limit on our cataphora resolution strategy given perfect parse, gender, and constraint information.

**Possessive Pronoun Antireflexive Filter Performance**

With only a handful of the 700 cases in the test set using our possessive pronoun antireflexive filters, we needed a larger test set to better assess performance. Thus we ran our algorithm on the entire

Table 6.4: Results of Extended Possessive Filter Test.

| Blocks | Correct | Incorrect | Rate |
|---|---|---|---|
| Verb | 4 | 3 | 57% |
| Prep | 37 | 24 | 61% |
| Total | 41 | 27 | 60% |

San Jose Mercury news corpus, 9246 identified cataphora cases in total, and checked every instance where the cataphora filter was applied to gauge its effectiveness (Table 6.4).

As Table 6.4 shows, of the 68 times the filters were applied, only 60% were needed – the other 40% were, in fact, acceptable cataphora. This performance is significantly worse than our anaphoric antireflexive filter results in Chapter 5. It will undoubtedly improve by collecting the filter patterns on a larger data set and with machine-learned rather than arbitrary thresholds. Furthermore, it should again be emphasized that blocking the patterns with the filter is much better than incorrectly linking the pronoun to the subject of the main clause. If an incorrect cataphora coreference is accepted, that resolution is incorrect and will not be changed. In theory, however, if one is incorrectly blocked, it will then be handled by the fall-back anaphora resolution procedure, which could quite likely still find the correct antecedent in a previous sentence, rather than cataphorically in the current one.

As well, the infrequent applicability of these filters is to be expected. We extracted them from a large but not overwhelmingly huge corpus, and applied stringent requirements for when they would subsequently be enforced. Future extraction should encompass more text, providing more filters, of higher trustworthiness.

### 6.2.7 Conclusion

Although they are not as frequent as anaphora, cataphora occur regularly in news articles, text databases, and in this very sentence. We are the first to address cataphora resolution as a separate procedure, to provide a systematic means of identifying and resolving cataphora, and to state performance results for cataphora resolution on a large data set. We showed intrasentential cataphora are the most common variety, and that resolving these instances in news articles can be done with a fully-automated system. In fact, cataphora resolution, which up until now was rarely and reluctantly handled by researchers, has demonstrated performance levels beyond anaphora resolution systems (compare to our 73% in Chapter 7).

Improvements to the cataphora resolution performance should be possible. Although we said we wished to develop a procedure that didn't use contextual information from previous sentences in the resolution, that information might be useful after all. If pronouns of the same gender occur in the previous sentence, for example, perhaps the pronoun in the cataphoric fragment can be linked safely to this other pronoun. A number of features might be considered when deciding between cataphora or anaphora resolution, and machine learning can be used to induce the optimum decisions.

# Chapter 7

# Unsupervised Anaphora Resolution

## 7.1 Introduction

This chapter is based on the Expectation Maximization approach to pronoun resolution [12]. In previous chapters, we presented unsupervised approaches to extracting useful coreference information, including probabilistic noun gender (Chapter 4), and antireflexive relations (Chapter 5). We now outline an approach that both extracts useful information while simultaneously resolving the pronouns in the corpus. This is done via Expectation Maximization.

Expectation Maximization is an iterative process. We begin with a number of pronouns to be resolved, and a set of candidate antecedents for each pronoun. Initially, each candidate noun is considered to be equally likely as the antecedent for the pronoun. Based on the co-occurrence of pronouns and nouns in this set-up, we can build probability distributions of noun gender, noun language model, noun antecedent likelihood, and distance from noun antecedent to pronoun. We thereby build probability distributions over individual nouns. Once we have these distributions, we can re-calculate the likelihood of each candidate noun in each pronoun candidate list according to a probability model based on these distributions. From the new likelihoods derived from this probability model, we can determine new counts, leading to new distributions, new likelihoods, etc., in an iterative fashion.

We perform this process for nouns in the AQUAINT Question Answering corpus [73]. Our major contribution is showing that unsupervised learning of anaphora resolution is possible. We can achieve results comparable to supervised methods (which, unlike our method, require training data). Further performance improvements can be obtained by providing an initializer for our probability distributions, and re-weighting the components of our probability model as feature functions in a log-linear model using maximum entropy.

As mentioned above, the majority of pronoun resolution approaches have thus far relied on manual intervention in the resolution process, such as using a manually-parsed corpus, or manually removing difficult non-anaphoric cases. Because of the modules developed in Chapter 6, we are now in a position to follow Mitkov et al.'s approach [53] with a fully-automatic pronoun resolution

method. Parsing, noun-phrase identification, and non-anaphoric pronoun removal can all be done automatically. Our algorithm is able to resolve *all* pronouns in our test documents. Furthermore, since we remove a large portion of the non-anaphoric pronouns in our training set as a pre-processing step, better data is fed to our EM learner.

As explained in Chapter 6, if we allow non-anaphoric pronouns to occur as instances to our training algorithm, non-sensical resolutions are made, polluting our noun probability distributions. Using the systems from Chapter 6 to both filter pleonastics and include future nouns when a cataphoric construction is detected therefore allow for cleaner distributions to be learned. Unfortunately, the unavoidable errors from pronouns referencing implicit phrases do reach our learner and also lower our final test performance results.

To the best of our knowledge, the only other fully unsupervised approach to reference resolution is by Cardie and Wagstaff [11], described in Chapter 2. Note that their approach is fundamentally different from our own in that their method relies on clustering, a process specific to individual documents, while our approach learns general probability models from an entire corpus.

Also of note for their use of EM for coreference resolution is the work of Ng and Cardie [61]. In this work, EM is used as an alternative to co-training. Classification accuracy of a supervised coreference resolution system can be modestly improved by using expectation maximization learning on the decisions of this superivsed classifier on a set of unlabelled data. One key difference from our system is that their learning does not model individual words, but simply tries to obtain a better combination of features than would be possible from the labelled set alone.

This work is also part of the tradition of unsupervised information extraction and bootstrapping techniques which we discussed in Chapter 2 and which we followed in other sections of this thesis. As discussed, these systems use the common assumption that there is a wealth of information in unlabelled data, and these riches can be extracted by assuming coreference links where possible and subsequently building probability distributions from the aggregate information. This chapter is unique, however, in that it begins by assuming every antecedent in the candidate list is equally likely, builds fractional counts of gender and other probability models from these assumed links, and then refines the counts and the antecedent selection process iteratively.

This kind of unsupervised method has its origin in the use of EM in bilingual word alignment. The prominent statistical methods in word alignment are all completely unsupervised, and our unsupervised method in particular is indebted to IBM's Model 1 [10]. The similarity can be seen most clearly if one considers our candidate list to be like the generating sentence, with the pronoun as the entity being generated. Melamed's competitive linking [49] method is also similar in its use of corpus-wide co-occurrence statistics and its use of a bag-of-words model within sentences. We also borrow one other technique from the statistical machine translation process: we will use the Maximum Entropy model weighting techniques from [62].

| | | | | | | |
|---|---|---|---|---|---|---|
| his: | $g$ | = | *masc* | $h$ | = | $g$'s family |
| | $C$ | = | arena (0), president (1) | | | |
| he: | $g$ | = | *masc* | $h$ | = | serenade $g$ |
| | $C$ | = | family (0), *masc* (1), arena (2), president (3) | | | |

Figure 7.1: Pronoun resolution instances from Sentence 1.

## 7.2 Resolution Framework

We use the formulation from [12], but with slightly different terminology. An instance of a pronoun to be resolved is regarded as a "triple," consisting of the pronoun's gender/number category, ($g$), the pronoun's parent, or syntactic head in the dependency tree, ($h$), and the candidate list of possible nominal antecedents, ($C$), creating a $(g, h, C)$ triple. Each candidate, $c$, in the candidate list, $C$, is itself a pairing consisting of its word value, $w$, the lexical string value of the candidate noun, and its jump value, $j$, representing the distance, in terms of intervening candidates, between that particular candidate and the pronoun; $c = (w, j)$.

As an example, recall Sentence 5 in Chapter 1, repeated here as Sentence 1:

(1) When the *president* entered the arena with *his* family, *he* was serenaded by a mariachi band.

Our candidate list construction process would convert this sentence into the two triples shown in Figure 7.1. The $j$-values follow each lexical candidate.

To prevent unlikely entities from being considered, and to increase the likelihood of ultimately resolving $g$ to the correct antecedent in $C$, we can immediately remove some candidates from consideration.

First of all, we restrict candidates to be preceding nouns occurring either in the current or previous sentence (except in cases of cataphora where future nouns may be added). Note, however, there is no fundamental restriction preventing candidates from being phrases, clauses, or other entities.

We also filter preceding nouns that have explicit gender in the text (see Chapter 4) that does not match the pronoun's gender. As we have seen, Binding Theory also provides constraints for candidate antecedents [23], and these are used in our system. Principle A in particular allows us to immediately select one antecedent for the pronoun, hence we only provide this single candidate unambiguously on the candidate list. We shall see that these unambiguous cases are useful for initializing the gender model (Section 7.4). A number of other constraints are outlined in [12].

To further improve the quality of information learned from our framework, we exclude pronoun triples from the learning process if the pronoun occurs in a sentences containing quotation marks. Entities in quotations are less likely to corefer with entities outside of quotations, and pronouns and pronoun candidates involving quotations can be challenging for anaphora resolution systems to resolve [35]. We avoid these issues in our learning framework by simply excluding them. The

quality of the individual instances provided to our learner thus increases. Unfortunately, for a fixed corpus side, removing some training instances reduces the overall amount of data being trained on. More data can be provided to offset these exclusions. Finally, note that in testing, we state results both on the non-quote portion of pronouns, and on all pronouns, where, as expected, we achieve inferior results (Section 7.7).

## 7.3 Expectation Maximization Anaphora Resolution Learning

Our ultimate aim is to learn probability distributions over the entities defined in the previous section, and to use these distributions for anaphora resolution. As explained in the introduction, this will be done by refining the proportions in a probability model iteratively using Expectation Maximization (EM) [16].

EM iteratively re-estimates probabilities in order to maximize the probability of a data set. We define the probability of our data set as the joint probability of all pronouns and their syntactic heads:

$$\Pr(Set) = \prod_{(g,h) \in Set} \Pr(g,h) \tag{7.1}$$

Ultimately, though, pronouns are symbolic references to real-world entities, and models for their behaviour are more informative if they depend on the antecedent to which the pronoun refers. We provide this dependency by incorporating the antecedent candidates in our probability model, allowing us to derive distributions over individual words. We expand each joint pronoun-parent probability over all the candidate antecedents for that pronoun:

$$\Pr(g,h) = \sum_{c \in C} \Pr(g,h,c) \tag{7.2}$$

We re-write the new joint probability as a conditional probability dependent on the particular candidate:

$$\Pr(g,h,c) = \Pr(g,h|c)\Pr(c) \tag{7.3}$$

This provides the full data set probability:

$$\Pr(Set) = \prod_{(g,h) \in Set} \sum_{c \in C} \Pr(g,h|c)\Pr(c) \tag{7.4}$$

Furthermore, we make the reasonable assumption that given a candidate, the pronoun and it's syntactic head are conditionally independent:

$$\Pr(g,h|c) = \Pr(g|c)\Pr(h|c) \tag{7.5}$$

Now we substitute the $c = (w, j)$ pairing for the $c$ values, and use the facts that $w$ and $j$ are independent and the gender and parent probabilities will naturally depend on $w$ but not $j$:

$$\Pr(g, h, c) = \Pr(g|w)\Pr(h|w)\Pr(w)\Pr(j) \tag{7.6}$$

The full data set probability is now:

$$\Pr(Set) = \prod_{(g,h) \in Set} \sum_{c \in C} \Pr(g|w)\Pr(h|w)\Pr(w)\Pr(j) \tag{7.7}$$

We iteratively refine the constituent probability models to maximize this full data set probability. There are four constituent distributions in this model. The $\Pr(g|w)$ distribution models the probability a particular word will be referenced by a pronoun of a particular gender/number. This is equivalent to the gender models we derived in Chapter 4. $\Pr(h|w)$ is the probability a particular candidate can have the same syntactic head as the pronoun. $\Pr(w)$ is the probability of a particular candidate occurring as an antecedent. As EM progresses, this factor becomes higher for words frequently occurring as antecedents, and diminishes for words rarerly referenced by pronouns. $\Pr(j)$ is the probability of an antecedent occurring at the distance of the current candidate. As EM progresses, this distribution approximates the recency criterion used in many anaphora resolution systems: candidates occurring closest to the pronoun are most likely to be antecedents, and this likelihood diminishes as distance increases.

There are two steps in EM, the E-step, where the likelihood of candidates is estimated from the probabilities distributions, and the M-step, where candidate likelihoods, taken as fractional counts, are used to re-estimate the probability distributions. Each requires information from the other, starting from some initial value. We explain them in order.

In the E-step, we are given values for the four constituent models. From these, we can define the likelihood of each candidate in the candidate list for a given $g$ and $h$, i.e., $\Pr(c|g, h)$, by dividing Equation 7.6 by Equation 7.2.

$$\Pr(c|g, h) = \frac{\Pr(g, h, c)}{\Pr(g, h)} \tag{7.8}$$

$$= \frac{\Pr(g|w)\Pr(h|w)\Pr(w)\Pr(j)}{\sum_{c' \in C} \Pr(g|w')\Pr(h|w')\Pr(w')\Pr(j')} \tag{7.9}$$

By looking at the $\Pr(c|g, h)$ values for all the candidates in a candidate list, we can rank the candidates in terms of their likelihood of being the antecedent. To test our anaphora resolution performance, we can resolve the pronoun to the highest-ranked candidate in this list.

For EM, the $\Pr(c|g, h)$ values are treated as fractional counts of the co-occurrence of the candidate with the pronoun and syntactic head. That is, when calculating the four constituent probabilities in the M-step, we count $c$ co-occurring with $(g, h)$ in the proportions specified by Equation 7.8.

For example, let $N(.)$ be the number of times a given event was observed. We can use our fractional counts of $\Pr(c|g,h)$ to provide us with $N(g,w)$ – the number of times we observe a particular noun with a particular gender, and $N(w)$, the number of times that noun occurred. Using these counts, we can re-estimate the gender probability as:

$$\Pr(g|w) = \frac{N(g,w)}{N(w)} \tag{7.10}$$

These two steps are repeated a number of times, until we are satisfied. In our work, we determine the number of iterations to run our algorithm by seeing how many repetitions result in the highest scoring models on the development key (Section 7.6). This number was found to be two; we use the models produced by the second M-step in our testing below.

## 7.4 Gender Model Initializer

In our experiments, as $(g,h,C)$ triples are produced for EM, roughly 9% of the instances of candidate lists are produced containing only one candidate. These unambiguous cases can be produced by Principle A reflexive pronoun resolutions, cataphora detection, through the filtering of all candidates but one using our other constraints, or simply because only this single noun precedes the pronoun in the current and previous sentence. Whatever the cause, when these cases arise, resolutions are automatic, and very high-quality information is counted by EM in building the probability distributions.

We investigated whether we could improve the models learned by our EM algorithm by initializing the iterative process with information from these unambiguous distributions. Ultimately, we found that beginning with an initial gender/number model for $\Pr(g|w)$ derived from the unambiguous cases resulted in the highest ultimate performance. That is, in the initial E-step, we weight candidates according to their unambiguous gender model probabilities, $\Pr_U(g|w)$. This unambiguous gender model is more sparse than the models learned through EM, as they are derived from an order of magnitude fewer instances. However, they are more accurate for the cases they do cover, arising as they do from unambiguous data.

An initializer is useful to EM for a number of reasons. First of all, there are a number of points of local optima where an EM algorithm might converge. Starting our algorithm with a gender/number model which is likely to be close to the desired gender/number distribution helps EM converge toward a better end state.

Since not every candidate is covered by the sparse unambiguous model, we use add-1 smoothing [30] to re-distribute the gender probability to unseen cases:

$$\Pr_U(g|w) = \frac{N_U(g,w)+1}{N_U(w)+4} \tag{7.11}$$

## 7.5 Maximum Entropy Formulation

Equation 7.8 shows that the selection of the antecedent for each pronoun involves choosing the most likely candidate based on the four constituent distributions. Since the denominator is constant for each candidate, we can re-write this selection for the antecedent $a$ as:

$$a = \text{argmax}_{c \in C} \Pr(c|g,h) \tag{7.12}$$

$$= \text{argmax}_{c=(w,j) \in C} \alpha \Pr(g|w)\Pr(h|w)\Pr(w)\Pr(j) \tag{7.13}$$

$$= \text{argmax}_{c=(w,j) \in C} \Pr(g|w)\Pr(h|w)\Pr(w)\Pr(j) \tag{7.14}$$

In this formulation, the antecedent is chosen based on the product of the four constituent distributions. It is possible that this is not the optimal combination for anaphora resolution accuracy. In the word-alignment models used in [62], Maximum Entropy is used to combine the constituent distributions as features in a log-linear model. Adopting this approach for our model, we can view $\Pr(g,h,c)$, that is $\Pr(g|w)\Pr(h|w)\Pr(w)\Pr(j)$ by Equation 7.6, as:

$$\exp\left(\begin{array}{c}\lambda_1 \log \Pr(g|w) + \lambda_2 \log \Pr(h|w) + \\ \lambda_3 \log \Pr(w) + \lambda_4 \log \Pr(j)\end{array}\right) \tag{7.15}$$

The new model introduces four new parameters, $\lambda_i, i = 1..4$. These parameters are set using the Maximum Entropy principle [6] to optimize the log-likelihood of a labelled training set. In our work, we set our weights using the limited memory variable metric method from Malouf's Maximum Entropy package [45], maximizing entropy on our antecedent-labelled development key (Section 7.6). Thus a small amount of labelled data is needed to set the parameters of the Maximum Entropy model; when we use this data our system is no longer completely unsupervised.

After setting the weights, we can use the log-linear model to select the antecedent from the candidate list:

$$a = \text{argmax}_{c \in C} \Pr(c|g,h) \tag{7.16}$$

$$= \text{argmax}_{c=(w,j) \in C} \exp\left(\begin{array}{c}\lambda_1 \log \Pr(g|w) + \lambda_2 \log \Pr(h|w) + \\ \lambda_3 \log \Pr(w) + \lambda_4 \log \Pr(j)\end{array}\right) \tag{7.17}$$

The relative size of the weights roughly corresponds to the importance of the individual constituent distributions. We provide the weights learned in our system in Table 7.1 below.

## 7.6 Experimental Set-up

We run our unsupervised learning process on two portions of the AQUAINT corpus, a portion used in development, and a portion used to produce our final test results. For development, we feed triples from 31K documents to our EM learner, representing about 333K pronouns. For testing, our learning set comes from 50K documents, representing about 890K pronouns. Each set has a corresponding

key, containing a subset of pronouns labelled with their antecedent. These keys correspond to the labelled AQUAINT data described in Chapter 3. Recall that we learn on only the portion of pronouns coming from sentences without quotation marks. Of the 1209 test key pronouns, 892 come from non-quotation sentences. We state results on both the full test key, and the non-quote portion we train on. We handle the unseen words and pairs in the full portion using additive smoothing, described in [12].

As in previous chapters, we define accuracy in terms of the number of pronouns correctly resolved to their antecedent. However, since we now handle non-anaphoric cases, we must also include these in our evaluation scheme. Mitkov et al. [53] point out that evaluating fully-automatic pronoun resolution with *success rate*, the percentage of anaphoric pronouns correctly resolved, neglects errors caused by attempting to resolve non-anaphoric pronouns, or incorrectly classifying pronouns as non-anaphoric. Thus we adopt a measure similar to their *resolution etiquette*. Our overall success rate score includes all pronoun cases – counting whether each anaphoric or cataphoric pronoun is correctly resolved to its antecedent, and whether each pleonastic or non noun-referential pronoun is correctly identified as being pleonastic or non-noun referential (but note again that while we can handle pleonastics, we will fail to correctly identify all implicit or non-noun referential pronouns at this time). When stating our results, we determine whether the difference in scores of two systems is statistically significant using McNemar's test with 95% confidence.

## 7.7 Results

We compare six different systems to evaluate unsupervised learning of anaphora resolution.

**1. Most Recent Noun**

For this system, we choose the closest candidate in the candidate list. Note this system will score higher than the most recent noun strategy in Chapter 4 because the candidate lists have been filtered as described in Section 7.2. If we apply Most Recent Noun without candiate list filtering on this set, we get 28.1%

**2. Uninitialized Expectation Maxization**

This system is our pure unsupervised method without initialization. We begin with uniform probability distributions and take the most likely candidate (based on the learned probability models) as the antecedent after two iterations of EM.

**3. Only Unambiguous Gender Probabilities**

This system chooses as the antecedent the most likely candidate according to the unambiguous-case, one-candidate candidate list gender distribution described in Section 7.4.

**4. Expectation Maxization Initialized with Unambiguous Gender Probabilities**

Here, we apply EM starting with initial gender probabilities, derived from the unambiguous gender cases. We run EM for another two iterations, and choose the antecedent from the candidates in the usual manner.

Table 7.1: Maximum Entropy Feature Function Weighting.

| Component | $\Pr(g\|w)$ | $\Pr(h\|w)$ | $\Pr(w)$ | $\Pr(j)$ |
|---|---|---|---|---|
| Weight ($\lambda$) | 0.931 | 0.056 | 0.070 | 0.167 |

Table 7.2: Resolution Performance on No-Quotation Test Set (%).

| Method | Overall Performance | Normalized by Upper Bound |
|---|---|---|
| 1 Most Recent Noun | 39.9 | 46.0 |
| 2 Unitialized EM | 63.2 | 72.8 |
| 3 Unambiguous Gender Only | 64.2 | 74.0 |
| 4 Gender-Initialized EM | 66.3 | 76.4 |
| 5 Maximum Entropy Model | 69.6 | 80.2 |
| 6 Upper Limit | 86.8 | 100.0 |

**5. Maximum Entropy Model**

This system builds on the probability distributions gathered from EM initialized with the unambiguous gender probabilities (number 4. above). Instead of multiplying the distributions according to the probability model, we use them as features in a log-linear model, as described in Section 7.5.

The weights of the log-linear model are learned using Maximum Entropy, and are provided in Table 7.1. Note that these weights essentially use gender as the determining factor, with much smaller weights for the other components. When two entities are equally likely in gender, however, the other components of the model, the distance, language model, and antecedent frequency, are incorporated. So, for example, if we're resolving the pronoun in "ate it." We can choose between "farmer" as an antecedent and "banana" based on the neutral gender of "banana." But if our decision lies between "coffee maker" and "banana," and both are neutral, we can turn to the language model to know that bananas can be eaten while coffee makers are not.

Recall that there are essentially two parts of our anaphora resolution process: building the candidate lists, and then selecting the most likely candidate according to the learned distributions. We sought a way to isolate and measure the quality of these two components separately. To do this, we determine the number of candidate lists where a true antecedent is available for resolution. We call this the "Upper Limit" on resolution accuracy according to the distributions. The value of the Upper Limit tells us the amount of error caused by building and filtering the candidate lists. Also, by normalizing our anaphora resolution score by this upper limit, we get the proportion of successful resolutions of the cases that were possible – telling us effectively how useful our distributions are for anaphora resolution. Ideally, we would want to get as high an Upper Limit as possible with as high a proportion of successes on the possible cases as possible.

Table 7.2 provides our anaphora resolution performance, with and without Upper Limit normalization, on the portion of the test set without quotations (that is, the proportion on which we run

Table 7.3: Resolution Performance on All Pronous, Pronouns from Sentences with Quotations

| Method | All | All Normalized | Quotes | Quotes Normalized |
|---|---|---|---|---|
| 1 Previous noun | 39.7 | 47.4 | 39.1 | 51.9 |
| 2 EM, no initializer | 61.0 | 72.8 | 54.9 | 72.8 |
| 3 Initializer, no EM | 62.8 | 74.9 | 58.7 | 77.9 |
| 4 EM w/ initializer | 63.2 | 75.4 | 54.6 | 72.4 |
| 5 Maximum Entropy Model | 66.9 | 79.8 | 59.3 | 78.6 |
| 6 Upper bound | 83.8 | 100.0 | 75.4 | 100.0 |

EM). Clearly, unsupervised learning of anaphora resolution is possible; we can reach an accuracy of over 63% using pure, unitialized EM on the candidate lists (23% higher than the most recent noun system). Perhaps surprisingly, even higher peformance is possible when only using the unambiguous gender information, 64.2%. This is surpassed by combining the gender initializer with EM, for 66.3% performance, for a statistically significant gain over un-initialized EM. Finally, re-weighting the learned EM models using Maximum Entropy and a log-linear model gives us our highest performance, 69.6%. This model is getting over 80% of the cases where a correct antecedent is possible (i.e., normalized by the upper limit), clearly showing the power of our learned distributions.

Although our results are most meaningful on the non-quotation portion on which we train, we wanted to test the performance of our system on all pronouns in text, including those from sentences with quotation marks. Results for both all pronouns and the added quotation-sentence subset are given in Table 7.3.

First of all, note the low upper limit on the quotation-sentence portion, 75.4%. Only about three-quarters of these pronouns have an antecedent in the canidate list. Observing this low proportion validates our decision to exclude these pronouns from the learning process (Section 7.2): the triples in these cases would be much noisier and pollute our learned distributions. Largely due to this low upper limit, all of the quotation-portion resolution systems score lower than the non-quotation portion on which we train.

Secondly, note that of the EM-systems, only the Maximum Entropy-enhanced system exceeds the gender-initializer system on the quotation-portion, and even then not significantly. There is thus little evidence that models learned with EM are useful on this unseen portion of instances. A better system for producing the quotation-pronoun candidate lists is clearly needed. These pronouns could then be included in the training phase of our algorithm, and allow for higher resolution accuracy on the "all pronoun" task.

When normalized by the upper limit, all of the Maximum Entropy systems score around 80% accuracy, yet this should not be regarded as an upper limit on resolution accuracy using our probability model. These models were themselves learned on a noisy set where only about 86.8% of pronouns had an antecedent in their candidate list. The learning process is polluted when EM attempts to learn something (add data to the distributions) from these impossible-to-resolve instances. Thus there is a

Table 7.4: Comparison to Chapter 4 System.

| Method | Accuracy |
|---|---|
| EM with gender initializer | 66.4 |
| Maximum Entropy Model | 70.8 |
| Chapter 4 System | 71.4 |

two-fold significance to the upper limit in our test results: it pollutes the learned data and reduces the number of cases that can be resolved. Thus understanding why and when an antecedent is missed in the candidate lists is vital to improving our system.

196 candidate lists in the "All Pronouns" task did not contain a true antecedent. Parser errors contribute to all other errors, and specifically remove the sole antecedent in 4 cases (2%). 41 pronouns (21%) had antecedents occurring outside our limited candidate window (Section 7.2). Incorrect pleonastic detection accounted for another 61 errors (31%), counting both errors of commission and omission. One cataphora construction was missed, leading to a single error (0.5%), while non-noun referential pronouns resulted in 49 errors (25%). The remaining 40 errors (20%) were caused by our various filters. From these statistics, it seems that better non-anaphoric pronoun handling should be an important future goal of our research.

### 7.7.1 Comparison to Chapter 4 System

In Chapter 4, we detailed a machine-learned anaphora resolution system that gathers gender information automatically from parsed corpora and the world wide web. This supervised system provides an interesting opportunity for comparison with our unsupervised approach, which also acquires gender information automatically. Most machine-learned anaphora resolution systems are supervised, and thus we can see whether our unsupervised approach is able to compete with the standard methodology.

We train the Chapter 4 SVM system on the full ANC training set of 1398 pronouns. Since the SVM system cannot handle pleonastic, implicit or cataphoric pronouns, these are removed from the test data for the comparison (slightly increasing the results over their Table 7.3 values), while quote-sentence pronouns were included since these are not handled in any special way by the SVM system. Results demonstrate our Maximum Entropy-enhanced system is quite competitive with the fully supervised SVM system, while the fully unsupervised approach is about 5% worse (Table 7.4). These results again provide encouragement for further development of the unsupervised methodology.

### 7.7.2 Top-$n$ Answers

Our system ranks all candidates in terms of their likelihood to be a pronoun's antecedent. We test this ranking by considering more than just the highest-ranked candidate. In this experiment, if any of the top-$n$ ranked candidates are valid antecedents, the case is considered correct. We compare

Table 7.5: Top-$n$ accuracy on "All" for $n = 1 \ldots 3$.

| Method | 1 | 2 | 3 |
|---|---|---|---|
| Previous noun | 39.7 | 61.4 | 72.1 |
| EM w/ initializer | 63.2 | 73.0 | 77.1 |
| Maximum Entropy Model | 66.9 | 76.4 | 80.3 |
| Maximum Entropy Model Normalized by Upper Limit | 79.8 | 91.2 | 95.8 |

our best EM model and our maxent extension to a heuristic that picks the $n$ candidates closest to the pronoun (Table 7.5). Both EM-based solutions provide better top-$n$ lists than the previous noun heuristic. Recall that the upper bound is only 83.8.

## 7.8 Conclusion

In this chapter we have explored the possibility of learning pronoun resolution systems from unlabelled data. This unsupervised approach allows for the training of powerful probability models that express the likelihood of gender/number, antecedent frequency, distance, and language modelling over individual words. Our initialized, yet fully unsupervised approach reaches a performance of 66% on the non-quote portion of the test set, while our Maximum Entropy extension, which uses a small amount of labelled training data, reaches 70%.

As mentioned earlier, we should next focus on better non-anaphoric pronoun handling, resulting in a higher upper limit on our candidate list resolution performance. Also, we may try incorporating more constituents into our probability model, or more features into the Maximum Entropy log-linear extension. We have yet to leverage whether a candidate has subject emphasis, occurs frequently in the document, or has occurred as an antecedent frequently elsewhere in a document. These features have all shown to improve performance in other systems.

# Chapter 8

# Conclusion

## 8.1 Contributions

In this thesis, we have explored machine learning from labelled and unlabelled corpora as a means to improve automated anaphora resolution. We have shown several ways that assumed coreference links in text can be used to bootstrap acquisition of useful statistical information.

We first looked at using a number of pre-defined patterns to acquire probabilisitic gender information, and showed 10% performance improvement using this data. We also showed how some difficult anaphora resolution cases can be resolved with new constraints learned from likely-coreferent or non-coreferent paths in text. Finally, we showed coreference assumptions and probabilistic information can be learned simultaneously in an iterative approach to unsupervised anaphora resolution.

This thesis has also reported the development of tools needed for fully-automatic anaphora resolution on unrestricted texts. Thousands of pronouns were manually labelled, enabling supervised learning and testing of coreference classifiers. We developed both pleonastic pronoun detection and cataphoric pronoun resolution, and we engineered our systems for the challenging domain of news articles, where a large vocabulary and a multitude of sentence styles are in use. The result of these labours has been our ability in Chapter 7 to state results for every pronoun in a number of news documents, rather than restricting our systems to special cases.

In summary, we now have a fully-automated anaphora resolution platform within which we can test future techniques and enhancements, and a large amount of labelled data to validate or refute our research ideas.

Although not yet prevalent, now is the time for fully-automated systems to become the standard in anaphora resolution. As researchers a decade ago embraced empirical testing over manual simulation of anaphora resolution algorithms, so must researchers of today embrace development on the full set of pronoun cases, with fully-automatic pre-processing modules devoid of any manual involvement or intervention. Otherwise our systems, so essential to modern information retrieval and text analysis, will nevertheless be irrelevant, incapable of performing on real applications with real data.

## 8.2 Future Work

There are three distinct directions for the future of our anaphora resolution research. The first is the expert-engineer direction. The second is the path of discovery. And the third is the path of application.

The expert-engineer direction works to solve the anaphora resolution problem by a series of clever enhancements. As the systems detailed in this thesis were developed, various new features and ideas for anaphora resolution began to pop-up every day. Naturally, the more time one spends working on a system, and the more familiar one becomes with the process, the more likely one is to identify interesting things to try. "Perhaps we might leverage the text formatting more," a voice said one morning. "Maybe coreference is less likely across paragraph breaks." Another day, the idea arose, "maybe we could use a key-phrase ranking algorithm to identify the key terms in the document, and then provide the noun rank as a feature in the resolution process – perhaps anaphora prefer more *important* antecedents?" "Maybe instead of resolving the pronouns linearly, we could search for the optimum order of resolutions to maximize overall antecedent probability?" These ideas are just a sample; the list of ideas to try grows larger and larger each day.

Any one of these ideas may in fact be an enhancement to the anaphora resolution process. The unfortunate aspect of this is, though, that on their own, each idea is unlikely to result in a statistically significant improvement in performance. And that is the name of the game – researchers make their livings by publishing things that are significant, otherwise, who is to say that the performance gain was not by chance? However, it is possible that anaphora resolution is nothing more than a series of a few thousand trivialities, special cases, and clever tricks, which individually are insignificant but together find the antecedent for every anaphor.

Is the path to solving anaphora resolution to try to write out routines for a diverse number of anaphora resolution situations? Should pronoun resolution be divided up into smaller and smaller sub-areas, perhaps with each research team choosing one specific pronoun, in one specific genre, and doing their best to encode the many sub-routines to handle the many ways this pronoun might link to its antecedent? Well, maybe – but we hope not!

Indeed, it seems to us to be more fruitful (and interesting) to try to automatically *discover* the clever tricks for resolving pronouns, without having to think them up on our own. This notion of *discovery* is different than previous notions of *learning*, because it implies more than feeding features and labels to a machine learning algorithm and letting this process decide how to combine them. Of course, supervised machine learning is extremely useful for this domain, as we have seen, but deciding what features to provide to the learner involves a degree of expert thinking and manual involvement on par with writing the rules in a rule-based system. And encoding all the rules or features needed to achieve top performance might turn out, as we speculated above, to be tricky.

So what is the path of discovery? How does it avoid manual involvement? It is hard to say now, but this thesis, and the antireflexive coreference relation chapter in particular (Chapter 5), seems to

be a step in the right direction. We showed how syntactic constraints can be extracted – what else can we gather in this manner? What rules can we learn – and how can we encode them? How do we prevent noise from polluting our results and consuming the infrequent cases that might be the key to solving the whole problem? These are difficult questions, but the answers don't seem entirely out of reach.

At the very least, we need to develop an evaluation methodology that not only tells us when we get something wrong, but lets us know how to fix it. For example, we use Binding Theory as a constraint or filter on candidate antecedents. Sometimes it fails and blocks an antecedent incorrectly, or includes one that should be blocked. Why, exactly, does this happen? Is it bad parses? What can we do about that? Is there a correlation between distance in sentences and Binding Theory applicability? In practice, does Binding Theory wear off with distance, or clause embedding, or some other quantifiable factor as Hobbs speculated [27]? We are in a position to answer these questions computationally – and automatically learn the clever tricks that will make the features we now have more powerful.

We mentioned three paths for future work and we have discussed expert rule development and ways to learn tricks automatically. The third path is the path of application. Anaphora resolution is an enabling technology; it's fun on its own but its ultimate worth is that it will assist other systems. Shouldn't we customize our anaphora resolution algorithms to suit their ultimate usage?

Natural Language Processing is indeed a subset of Artificial Intelligence, which implies that studies in the NLP domain should consider how an intelligent agent might make use of the tools we devise. One intelligent agent might be a Question Answering application. Users ask natural language queries and the agent determines the answers in gigabytes of text. We suggested earlier that developing an anaphora resolution application for this task would boost performance. Rather than just trying various anaphora resolution systems as pre-processing to QA, we should lay a foundation of theoretical justification for designing our anaphora resolution systems in a certain way.

For example, we might first investigate: given a standard set of test questions from the QA task at TREC, how many would benefit from anaphora resolution, and what kinds of resolution? Or conversely, we could identify the documents that a subset of questions come from, label them for coreference, and see the extent to which performance increases. Inspecting the errors and successes in these situations would be very instructive; running these tests might compel us to design the anaphora resolution process in a new way, with different kinds of preferences and biases.

In research, as in life, there are always trade-offs and compromises. The ultimate direction of this work will indubitably be a combination of the three directions we have sketched above, with new paths and branches of inquiry included that we cannot foresee at the moment. We can be confident, however, that the direction will be forward, that new possibilities will arise as new technologies develop, and that the demand for anaphora resolution will not diminish, but grow. In other words, it's an exciting time for this area.

# Bibliography


[1] José Abraços and José Gabriel Lopes. Extending DRT with a focusing mechanism for pronominal anaphora and ellipsis resolution. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 1128–1132, 1994.

[2] Chinatsu Aone and Scott William Bennett. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 122–129, 1995.

[3] Breck Baldwin. CogNIAC: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 38–45, 1997.

[4] David L. Bean and Ellen Riloff. Corpus-based identification of non-anaphoric noun phrases. In *ACL*, 1999.

[5] David L. Bean and Ellen Riloff. Unsupervised learning of contextual role knowledge for coreference resolution. In *HLT-NAACL*, pages 297–304, 2004.

[6] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[7] Shane Bergsma. Automatic acquisition of gender information for anaphora resolution. In *To appear: Proceedings of the Eighteenth Canadian Conference on Artificial Intelligence (AI'2005)*, 2005.

[8] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of COLT*, pages 92–100, 1998.

[9] Susan Brennan, Marilyn Friedman, and Carl Pollard. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162, 1987.

[10] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–312, 1993.

[11] Claire Cardie and Kiri Wagstaff. Noun phrase coreference as clustering. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 82–89, 1999.

[12] Colin Cherry and Shane Bergsma. An expectation maximization approach to pronoun resolution. In *[To Appear] Proceedings of the Ninth Conference on Natural Language Learning (CoNLL-2005)*, 2005.

[13] Noam Chomsky. *Lecutres on Government and Binding*. Foris Publications, Dordrecht, Holland, 1981.

[14] William Cohen. Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123, 1995.

[15] Ido Dagan and Alan Itai. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, volume 3, pages 330–332, Helsinki, Finland, 1990.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[17] Richard Evans. Applying machine learning toward an automatic classification of *it*. *Literary and Linguistic Computing*, 16(1):45–57, 2001.

[18] Richard Evans. Refined salience weighting and error analysis in anaphora resolution. In *Proceedings of Reference Resolution for Natural Language Procesing*, pages 51–59, 2002.

[19] Richard Evans and Constantin Orăsan. Improving anaphora resolution by identifying animate entities in texts. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference*, pages 154–162, 2000.

[20] David Fisher, Stephen Soderland, Joseph McCarthy, Fangfang Feng, and Wendy Lehnert. Description of the UMass system as used for muc-6. In *Proceedings of the Sixth Conference on Message Understanding (MUC-6)*, pages 127–140, 1995.

[21] Niyu Ge, John Hale, and Eugene Charniak. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–171, 1998.

[22] Barbara Grosz and Candice Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[23] Liliane Haegeman. *Introduction to Government & Binding theory: Second Edition*. Basil Blackwell, Cambridge, UK, 1994.

[24] Sanda Harabagiu, Razvan Bunescu, and Steven Maiorano. Text and knowledge mining for coreference resolution. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics*, pages 55–62, 2001.

[25] Sanda Harabagiu and Steven Maiorano. Knowledge-lean coreference resolution and its relation to textual cohesion and coherence. In *Proceedings of the ACL-99 Workshop on the Relation of Discourse/Dialogue Structure and Reference*, pages 29–38, 1999.

[26] Donna Harman. The DARPA TIPSTER project. *ACM SIGIR Forum*, 26(2):26–28, 1992.

[27] Jerry Hobbs. Resolving pronoun references. *Lingua*, 44(311):339–352, 1978.

[28] Nancy Ide and Keith Suderman. The american national corpus first release. In *Proceedings of the Fourth Language Resources and Evaluationi Conference*, pages 1681–84, 2004.

[29] Hideki Isozaki and Tsutomu Hirao. Japanese zero pronoun resolution based on ranking rules and machine learning. In *Proceedings of EMNLP-2003*, pages 184–191, 2003.

[30] Harold Jeffreys. *Theory of Probability*, chapter 3.23. Oxford: Clarendon Press, 3rd edition, 1961.

[31] Thorsten Joachims. Making large-scale SVM learning practical. In B. Schölkopf and C. Burges, editors, *Advances in Kernel Methods*. MIT-Press, 1999.

[32] D. Jurafsky and J. H. Martin. *Speech and language processing*. Prentice Hall, 2000.

[33] Megumi Kameyama. Recognizing referential links: An information extraction perspective. In *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 46–53, 1997.

[34] Andrew Kehler. Probabilistic coreference in information extraction. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 163–173, 1997.

[35] Christopher Kennedy and Branimir Boguraev. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 113–118, 1996.

[36] Beata Klebanov. Using latent semantic analysis for pronominal anaphora resolution. Master's thesis, University of Edinburgh, Edinburgh, 2001.

[37] R. Langacker. On pronominalization and the chain of command. In D. Reibel and S. Schane, editors, *Modern studies in English*, pages 160–186. Prentice-Hall, 1969.

[38] Shalom Lappin and Herbert J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.

[39] Dekang Lin. Principar - an efficient, broad-coverage, principle-based parser. In *Proceedings of COLING-94*, pages 42–48, Kyoto, Japan, 1994.

[40] Dekang Lin. University of Manitoba: Description of the PIE system as used for MUC-6. In *Proceedings of the Sixth Conference on Message Understanding (MUC-6)*, 1995.

[41] Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computation Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING/ACL-98)*, pages 768–773, 1998.

[42] Dekang Lin. Dependency-based evaluation of MINIPAR. In *Proceedings of the Workshop or the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*, 1998.

[43] Dekang Lin. LaTaT: Language and text analysis tools. In *Proceedings of the First International Conference on Human Language Technology Research*, 2001.

[44] Dekang Lin and Patrick Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360, 2001.

[45] Robert Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*, pages 49–55, 2002.

[46] Ronnie Mann and Catriona McPherson. Interclaual cataphora in English. *Unpublished*, 1999.

[47] Joseph McCarthy and Wendy Lehnert. Using decision trees for coreference resolution. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1050–1055, 1995.

[48] Joseph F. McCarthy. *A Trainable Approach to Coreference Resolution for Information Extraction*. PhD thesis, University of Massachusetts Amherst, 1996.

[49] I. Dan Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, 2000.

[50] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[51] Ruslan Mitkov. Factors in anaphora resolution: they are not the only things that matter. a case study based on two different approaches. In *Proceedings of the ACL '97 / EACL '97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, pages 14–21, 1997.

[52] Ruslan Mitkov. Anaphora resolution: The state of the art. Technical report, University of Wolverhampton, Wolverhampton, 1999.

[53] Ruslan Mitkov, Richard Evans, and Constantin Orasan. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 168–186, 2002.

[54] Ruslan Mitkov and Paul Schmidt. On the complexity of pronominal anaphora resolution in machine translation. In Carlos Martín-Vide, editor, *Mathematical and computational analysis of natural language*. John Benjamins Publishers, Amsterdam, 1998.

[55] MUC-6. Coreference task definition (v2.3, 8 Sep 95). In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 335–344, 1995.

[56] MUC-7. Coreference task definition (v3.0, 13 Jul 97). In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1997.

[57] Christoph Müller, Stefan Rapp, and Michael Strube. Applying co-training to reference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 352–359, 2002.

[58] Vincent Ng and Claire Cardie. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, pages 730–736, 2002.

[59] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, 2002.

[60] Vincent Ng and Claire Cardie. Bootstrapping coreference classifiers with multiple machine learning algorithms. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 113–120, 2003.

[61] Vincent Ng and Claire Cardie. Weakly supervised natural language learning without redundant views. In *Proceedings of the HLT-NAACL*, pages 94–101, 2003.

[62] Franz J. Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, PA, July 2002.

[63] Patrick Pantel and Deepak Ravichandran. Automatically labeling semantic classes. In *HLT-NAACL*, pages 321–328, 2004.

[64] J. Peral and A. Ferrández. Translation of pronominal anaphora between English and Spanish languages: discrepancies and evaluation. *Journal of Artificial Intelligence Research*, 18:117–147, 2003.

[65] Steven Pinker. *The Language Instinct*. William Morrow and Company, New York, 1994.

[66] Tony Rose, Mark Stevenson, and Miles Whitehead. The reuters corpus volume 1 – from yesterday's news to tomorrow's language resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 29–31, 2002.

[67] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: a modern approach*, chapter 20: Statistical Learning Methods, page 720. Prentice Hall, Upper Saddle River, N.J., 2nd edition edition, 2003.

[68] Pieter Seuren. Automatic summarization by paragraph initial sentences extraction. In *Rencontre Internationale sur l'extraction, le filtrage et le Résumé Automatique*, pages 64–71, 1998.

[69] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.

[70] Roland Stuckardt. Coreference-based summarization and question answering: a case for high precision anaphor resolution. In *Proceedings of the 2003 International Symposium on Reference Resolution and Its Application to Question Answering and Summarization (ARQAS)*, pages 33–41, 2003.

[71] J. Tetreault. Analysis of syntax-based pronoun resolution methods. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 602–605, 1999.

[72] Peter D. Turney. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4):303–336, 2000.

[73] Ellen Vorhees. Overview of the TREC 2002 question answering track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC)*, 2002.

[74] Dave Winer. Bootstrapping. http://davenet.scripting.com/2000/11/30/bootstrapping, 2000.